Final Research Documentation

Cancer Statistics in the United States

DHARTI RATHOD - 84213
KAVITHA PRIYADHARSINI RAVI - 85144
RANSEY BAUTISTA - 82176
SHRI RACHANA RAJASEKARAN - 85390

INTERNATIONAL TECHNOLOGICAL UNIVERSITY

CS 960 - Introduction to Data Science

Spring 2014

Document Version 1.00

Last Update: April 8, 2014

i. Preface

"We live in a universe devoted to the creation, and eradication, of awareness. Augustus Waters did not die after a lengthy battle with cancer. He died after a lengthy battle with human consciousness, a victim - as you will be - of the universe's need to make and unmake all that is possible (Green, 8) " .

— John Green, *The Fault in Our Stars*

"Data that is loved tends to survive."

— Kurt Bollacker, Data Scientist, *Freebase/Infochimps*

ii. Acknowledgements

Above all else, Data is fact and fact is important to base a theory or its projections upon. We are thankful to Professor Helen Huynh, for helping us understand the importance of Data science in the real world today. Also for her excellent guidance and a well structured and paced out classroom program to complete the project with a deep understanding of its underlying concepts. We are grateful to her for helping us pick this project, that would impact people's lives . An analysis about cancer through the data that we have done ,would bring about a great amount of awareness not only to the organization that is concerned with cancer awareness, but also to practice our knowledge on how to be a data scientist in the real world.

Data science is an emerging branch of Information Technology. This project will help us realize if this field is for us. The group would also like to thank the website http://www.cancer.org/ for providing us the data necessary for this project. Our sincerest thanks to everyone who has helped us in the process of understanding, working on it and hence completing our project . Last but not the least we thank for the guidance and support of our fellow classmates and our families. Special thanks goes to ITU, for providing such infrastructure needed for the term .

Revision History

Date	Version	Author	Description
02/09/2014	0.01	Ransey	The template of this document is created.
02/25/2014	0.02	Ransey	Preface, Acknowledgements and Introduction are written.
03/12/2014	0.03	Ransey	Theoretical is written.
03/12/2014	0.04	Dharti Kavitha	Hypothesis/Goal is stated.
03/13/2014	0.05	Rachana	Methodology is created.
03/13/2014	0.06	Ransey	Summary is started-up.
03/13/2014	0.07	Kavitha	Recognition/Comments is started-up.
03/13/2014	0.08	Dharti	Project Schedule is updated.
03/16/2014	0.09	Ransey	Recognition/Comments is revised.
03/19/2014	0.10	Ransey	Implementation is started.
03/24/2014	0.11	Dharti	Bibliography
03/25/2014	0.12	Ransey Kavitha Dharti	Implementation is revised.
03/25/2014	0.13	Rachana	Several parts document revision
04/06/2014	0.14	All	Last minute changes of some parts.
04/08/2014	1.00	All	Final document for submission.

iii. List of Figures

iv. List of Pivot Tables

v. List of Tables

vi. Table of Contents

1. Introduction

Data is everywhere , from our biological system to the digital footprint that we leave on the web. In todays era every data is considered valuable. Data Science field , is combination of data analysis , data mining , statistics , computer science , design and also include the basic understanding of the field from which the data originates. To work with data as a data scientist , one has to be inquisitive, creative and should have skills to analyse data from different angles, report the findings and embrace change whenever it is necessary. This report builds a bridge between the theory of data science and a practical institution of the same .

Our project report deals with cancer data in the United States of America from the year 2010 to 2014. To begin with, let us first understand what is cancer and how data science's principles and tools can be used to understand its prevalence in the united States. Our body is made up of trillions of cells. Cell is basic structural and functional unit of all living organism. The cells replicate independently allowing the person to grow and as we grow older, the divided cells replace worn out or injured. Thus cells are known as "building block of life". However, disease name Cancer, medically known as *Malignant neoplasia*, is a broad group of diseases involving unregulated cell growth. Cells divide and grow uncontrollably forming tumors which invades nearby or distant body parts, tissue of organs or skin and cause health issues and some lead to death if not treated early. There are over 200 different types of cancer that could affect us according to the data that we obtained from www.cancer.org. There are types of cancer that cannot be completely cured even if the treatment is started early .Cancer could be tumor causing however, leukemia is not known to form tumors as often as others. In this case, they move through the bloodstream of the person affecting other internal organs. Cancer cells are mutated cells that are tumor causing or invading other organs in the body and affecting the normal activity of that organ. Hence causing issues to the human body.

As we read this report , there are millions of people across the globe who have been affected by cancer , are susceptible to cancer due to inheritance of mutated genes or was once affected by cancer. This study shows data from the last few years about cancer and its types in different states of USA. As you proceed reading report, we hope to bring in a better understanding to how many people have been affected by cancer across different age groups and gender and in a specific state. The final conclusions helps a person decide the type of screening for cancer that they might choose to do more often than other. Also, this helps

medical health representatives be well prepared with the required resources to treat people in that years for the years to come . In helping the government deciding on the research funding needed for cancer centres in that states . We are hoping to help people from the government and private sector to understand the most prevalent type of cancer in a specific state to help improve patient's health and for NGO to create better awareness in that state for that cancer . Breaking down the type of cancer with respect to the type that affects a particular gender and the age group most susceptible to it , helps NGOs reach people faster and insurance providers to be aware of what coverage to include to make sure they are profitable and at the same time people who are subscribed have complete access to the medical facilities.

2. Theoretical

As a part of the team project, each team which consisted of three member was required to come up with three titles per person . We discussed about the tools that could be used, the data that is available and most importantly how the effect we would have on the society . The final four titles that we had on the table were -Studying Cancer data , Weather Trend, Mobile Traffic Applications and Million Song Data . The next step was for us, was to take a sample data and analyze to understand what we would achieve with the data we had.

- **Million Song Data**

We had the analysis heading towards analysing the different instruments, lyrics and rhythm used for the million songs from an archive to find the most prevalent rhythm used, most repeated words and the most commonly used instruments.This is to generalize a go-to rhythm and lyrics that would be far fetching . In the context of music, this would have samples of rhythm and a set of words most commonly used by songwriters and composers with greater success with their songs.

- **Weather Trend**

This included a decade's worth data about the weather that has been recorded across the globe, studying that would give us the results close to the amount of increase in greenhouse gases, the rate at which climate change is happening and the most susceptible places for hurricane, cyclone, tsunami and earthquake in the years to come.

- **Mobile Traffic analysis**

Data analysis for the mobile traffic that has been recorded in the last five years across the globe would give us an idea about how much we could be expecting it to increase in the next few years. However, moore's law does not apply to this as the number of devices are increasing exponentially and at the same time, some parts of the world are still not clearly connected to the rest of the world. Our study was to help the tech companies to predict the increase the mobile device sales and the amount of infrastructure needed to support the mobile devices and the traffic caused by them.

- **Cancer data analysis**

Analysis of the data for the last few years across the types of cancer prevalent in the united states. As we proceeded to try out a simple excel sheet analysis, we understood that the data provided was divided across the different states based on gender and age as well. The outcome is the most prevalent type in a state based on gender and age. This would help the government to decide on the funding they provide for

research and treatment of cancer in the states. It would greatly help the NGO's to educate people through awareness programs on the most prevalent in the state instead of trying to scare people with reaching about all the 200 types of cancer. Finally , the graphs could be easily printed or posts on social networks for people to know the risk zone they are in and get themselves checked regularly for that type. As this seemed to help a large number of people and the final results could be really helpful to make people understand the need for cancer screening , we chose cancer analysis as our project .

Problem Statements :

#1.Identify which state in the United States is at the maximum risk zone for cancer.

#2. Detect the most prevalent type of cancer in the coming years.

#3.Which cancer is deadliest for male and female in the US.

3. Hypothesis

We are doing statistical hypothesis test from an observational study using the data we have retrieved from www.cancer.org. The data consist of the number of new cases of patients affected by cancer every year for the last five years across all the states in the US, also the number of deaths in these states in those years.

Our goal is to prove our hypotheses enumerated below:

Positive:

1. California is the maximum risk zone state in the United States based on initial data.
2. Prostate cancer and Breast cancer will continue to be dominant in the coming years.
3. Lung & Bronchus cancer will be deadliest for both male and female in the US.

Negative:

1. California is not the maximum risk zone state if ratio by state population is considered in the data.

4. Methodology

How did we collect our data?

Data was extracted from the American Cancer Society's website "www.cancer.org" as .pdf documents.This data has to be processed, they had to be converted to excel files/ comma seperated files for accessing data withing these files. This processing was a must because of the pictures in the original pdf that we weren't interested in using in our project. Data warehousing was done to place all the related data together in our repository to do data analysis on that .

The predictions for the year 2014 has been provided in www.cancer.org however, we are using our data mining technique to do the same because our group wants our research and analysis to be cross verified for to be more realistic.

As stated, we used the 2014 cancer data as our data set to compare the results and we are using the 5 year data (from 2010-2014) that is also readily available in the website.

Method followed:

We used the deductive argument method while reasoning our hypotheses from the data that we massaged .

Technologies/Tools used?

For this project, we are not using a database. Since we took a lot of time deciding which data set to use, we didn't have time to make a schema for relational database to store our data, and settled for the MS Excel to do our data manipulation instead. We are using R to project the graphs from the data that we have obtained.

5. Implementation

For our implementation, we have created a flowchart that we have used in deriving an answer to each of our problems.

Problem #1 Identify which state in the United States is at the maximum risk zone for cancer.

Since our data gives "new cases" and "death cases" for cancer in each state in the United States, our group decided to use both data to show two sides of "maximum risk" of each year.

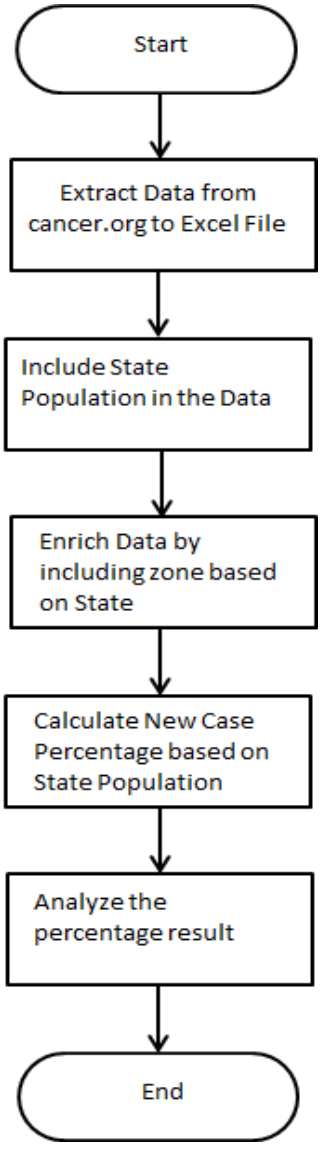


Figure 5.1 New Cases Process

The flowchart below shows the process for death cases by state, which is similar to Figure 5.1 New Case Process.

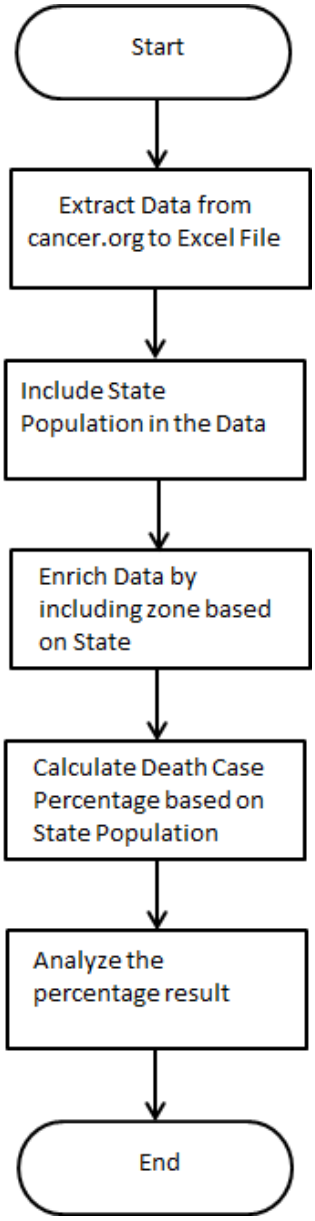


Figure 5.2 Death Cases Process

Part of both the processes includes Extraction of Data from cancer.org website.

Illustration of data transformation is shown below:

Table 5.1:Original Data from the website (extracted to excel file)

State	All Sites	Female Breast	Uterine Cervix	Colon & Rectum	Uterine Corpus	Leukemia	Lung & Bronchus	Melanoma-Skin	Non-Hodgkin Lymphoma	Prostate	Urinary Bladder
Alabama	26,770	3,660	210	2,350	650	690	4,160	1,320	990	3,760	990
Alaska	3,750	450	†	280	100	100	430	90	140	530	150
Arizona	32,830	4,520	210	2,560	910	950	4,280	1,430	1,320	4,390	1,490
Arkansas	16,520	2,050	140	1,500	400	480	2,660	490	660	2,240	640
California	171,730	26,130	1,550	13,930	5,650	5,650	18,780	8,440	7,770	23,010	7,210
Colorado	23,810	3,780	160	1,720	750	870	2,540	1,400	1,060	3,680	1,040
Connecticut	22,070	3,160	120	1,650	790	610	2,730	1,090	920	3,120	1,170

We added State Population to obtain the ratio instead of looking at the number of cases alone.

Table 5.2:State Population Added to Extracted Data

State	Population	All Sites	Female Breast	Uterine Cervix	Colon & Rectum	Uterine Corpus	Leukemia	Lung & Bronchus	Melanoma-Skin	Non-Hodgkin Lymphoma	Prostate	Urinary Bladder
Alabama	4,888,343	26,770	3,660	210	2,350	650	690	4,160	1,320	990	3,760	990
Alaska	760,935	3,750	450	†	280	100	100	430	90	140	530	150
Arizona	6,869,821	32,830	4,520	210	2,560	910	950	4,280	1,430	1,320	4,390	1,490
Arkansas	3,003,467	16,520	2,050	140	1,500	400	480	2,660	490	660	2,240	640
California	39,444,164	171,730	26,130	1,550	13,930	5,650	5,650	18,780	8,440	7,770	23,010	7,210
Colorado	5,519,141	23,810	3,780	160	1,720	750	870	2,540	1,400	1,060	3,680	1,040
Connecticut	3,618,375	22,070	3,160	120	1,650	790	610	2,730	1,090	920	3,120	1,170

We enrich this data by including the zone of each state based on time-zone it belongs to. The percentage (%) of "All Sites" cases is calculated based on the Population of the state.

Table 5.3:Percentage of "all sites" cases calculation

State	Zone	Population	All Sites	%	Female Breast	%FB	Uterine Cervix	%UCv	Colon & Rectum	%C/R	Uterine Corpus	%UCo
Alabama	Central	4,888,343	26,770	0.5476%	3,660	0.0749%	210	0.0043%	2,350	0.0481%	650	0.0133%
Alaska	Alaskan	760,935	3,750	0.4928%	450	0.0591%	†		280	0.0368%	100	0.0131%
Arizona	Mountain	6,869,821	32,830	0.4779%	4,520	0.0658%	210	0.0031%	2,560	0.0373%	910	0.0132%
Arkansas	Central	3,003,467	16,520	0.5500%	2,050	0.0683%	140	0.0047%	1,500	0.0499%	400	0.0133%
California	Pacific	39,444,164	171,730	0.4354%	26,130	0.0662%	1,550	0.0039%	13,930	0.0353%	5,650	0.0143%
Colorado	Mountain	5,519,141	23,810	0.4314%	3,780	0.0685%	160	0.0029%	1,720	0.0312%	750	0.0136%
Connecticut	Eastern	3,618,375	22,070	0.6099%	3,160	0.0873%	120	0.0033%	1,650	0.0456%	790	0.0218%

Final outcome was 5 years of data that has been processed as shown above in the illustrated steps to further conduct the needed analysis on this data. The data that has been processed are the new cases and death cases of the 5 years years of data that we had collected. This is further combined to provide the complete 5 years data (2010-2014). This result is explained in Part 6 of Analysis section.

Problem #2 Detect the most prevalent types of cancer in the coming years. For this problem, we used the "new cancer cases" data for the last 5 years.

For Men:

Table 5.4:New cancer cases in Men (2010-2014)

Cancer Type	2010 Male New Case	2011 Male New Case	2012 Male New Case	2013 Male New Case	2014 Male New Case	2010 Female New Case	2011 Female New Case	2012 Female New Case	2013 Female New Case	2014 Female New Case
Liver & intrahepatic bile duct	17,430	19,260	21370	22,720	24,600	6,690	6,930	7,350	7,920	8,590
Gallbladder & other biliary	4,450	3,990	4480	4,740	4,960	5,310	5,260	5,330	5,570	5,690
Pancreas	21,370	22,050	22090	22,740	23,530	21,770	21,980	21830	22,480	22,890
Other digestive organs	1,660	1,690	1800	1,900	1,880	3,220	3,310	3,890	3,850	3,880
Larynx	10,110	10,160	9840	9,680	10,000	2,610	2,580	2,520	2,580	2,630
Lung & bronchus	116,750	115,060	116470	118,080	116,000	105,770	106,070	109690	110,110	108,210
Other respiratory organs	3,740	3,670	3960	4,000	4,000	1,630	1,780	1700	1,760	1,710
Bones & joints	1,530	1,620	1600	1,680	1,680	1,120	1,190	1,290	1,330	1,340
Soft tissue (including heart)	5,680	6,050	6110	6,290	6,550	4,840	4,930	5,170	5,120	5,470
Melanoma-Skin	38,870	40,010	44250	45,060	43,890	29,260	30,220	32,000	31,630	32,210
Breast	1,970	2,140	2190	2,240	2,360	207,090	230,480	226,870	232,340	232,670
Uterine cervix						12,200	12,710	12,170	12,340	12,360
Uterine corpus						43,470	46,470	47130	49,560	52,630
Ovary						21,880	21,990	22,280	22,240	21,980
Vulva						3,900	4,340	4,490	4,700	4,850
Vagina & other genital, female						2,300	2,570	2680	2,890	3,170
Prostate	217,730	240,890	241740	238,590	233,000					
Testis	8,480	8,290	8590	7,920	8,820					
Penis & other genital, male	1,250	1,360	1570	1,570	1,640					
Urinary Bladder	52,760	52,020	55600	54,610	56,390	17,770	17,230	17910	17,960	18,300
Kidney & renal pelvis	35,370	37,120	40250	40,430	39,140	22,870	23,800	24520	24,720	24,780
Ureter & other urinary organs	1,490	1,610	1760	1,760	1,890	1,000	1,120	1,100	950	1,110

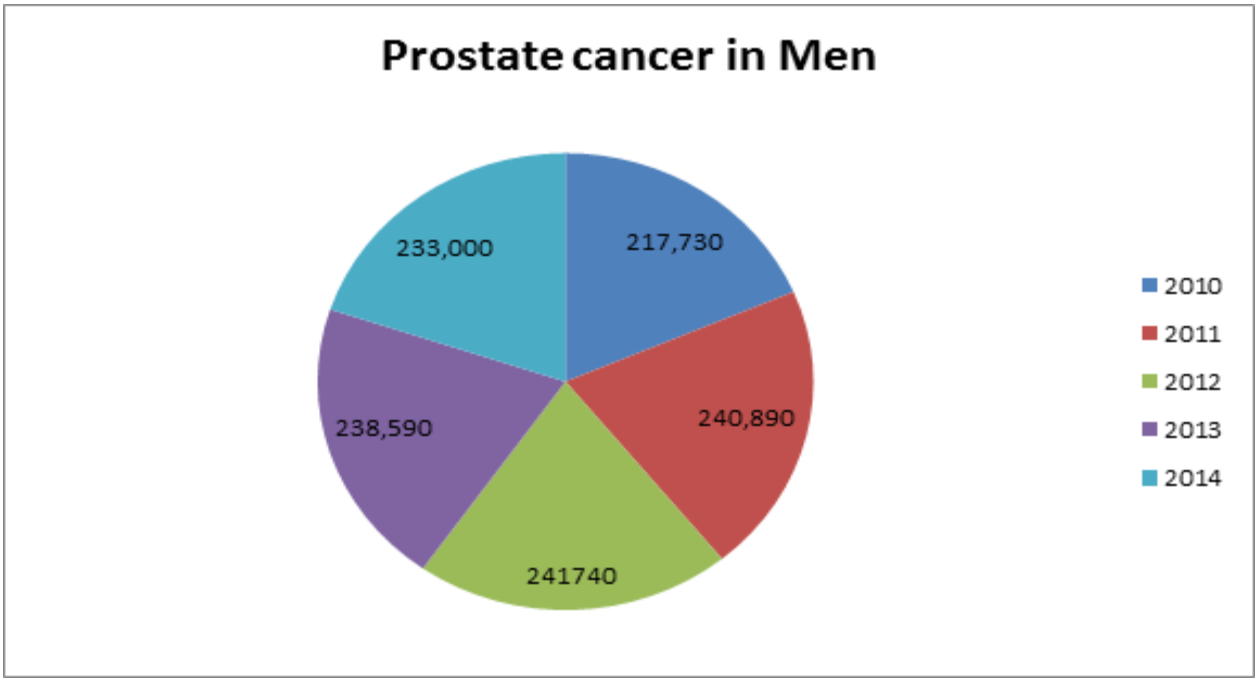


Figure 5.3 Prostate Cancer Cases (2010-2014)

For Women:

Table 5.5:New cancer cases in Women (2010-2014)

Cancer Type	2010 Male New Case	2011 Male New Case	2012 Male New Case	2013 Male New Case	2014 Male New Case	2010 Female New Case	2011 Female New Case	2012 Female New Case	2013 Female New Case	2014 Female New Case
Tongue	7,690	8,560	9040	9,900	9,720	3,300	3,500	3,730	3,690	3,870
Mouth	6,430	6,950	7,030	6,730	7,150	4,410	4,560	4,590	4,670	4,770
Pharynx	9,880	10,600	10790	11,200	11,550	2,780	2,980	2720	2,730	2,860
Other oral cavity	1,420	1,600	1680	1,790	1,800	630	650	670	670	720
Esophagus	13,130	13,450	13950	14,440	14,660	3,510	3,530	3510	3,550	3,510
Stomach	12,730	13,120	13020	13,230	13,730	8,270	8,400	8,300	8,370	8,490
Small intestine	3,680	3,990	4380	4,670	4,880	3,280	3,580	3690	4,140	4,280
Colon†	49,470	48,940	49920	50,090	48,450	53,430	52,400	53,250	52,390	48,380
Rectum	22,620	22,910	23500	23,590	23,380	17,050	16,960	16790	16,750	16,620
Anus, anal canal, & anorectum	2,000	2,140	2250	2,630	2,660	3,260	3,680	3980	4,430	4,550
Liver & intrahepatic bile duct	17,430	19,260	21370	22,720	24,600	6,690	6,930	7,350	7,920	8,590
Gallbladder & other biliary	4,450	3,990	4480	4,740	4,960	5,310	5,260	5,330	5,570	5,690
Pancreas	21,370	22,050	22090	22,740	23,530	21,770	21,980	21830	22,480	22,890
Other digestive organs	1,660	1,690	1800	1,900	1,880	3,220	3,310	3,890	3,850	3,880
Larynx	10,110	10,160	9840	9,680	10,000	2,610	2,580	2,520	2,580	2,630
Lung & bronchus	116,750	115,060	116470	118,080	116,000	105,770	106,070	109690	110,110	108,210
Other respiratory organs	3,740	3,670	3960	4,000	4,000	1,630	1,780	1700	1,760	1,710
Bones & joints	1,530	1,620	1600	1,680	1,680	1,120	1,190	1,290	1,330	1,340
Soft tissue (including heart)	5,680	6,050	6110	6,290	6,550	4,840	4,930	5,170	5,120	5,470
Melanoma-Skin	38,870	40,010	44250	45,060	43,890	29,260	30,220	32,000	31,630	32,210
Breast	1,970	2,140	2190	2,240	2,360	207,090	230,480	226,870	232,340	232,670

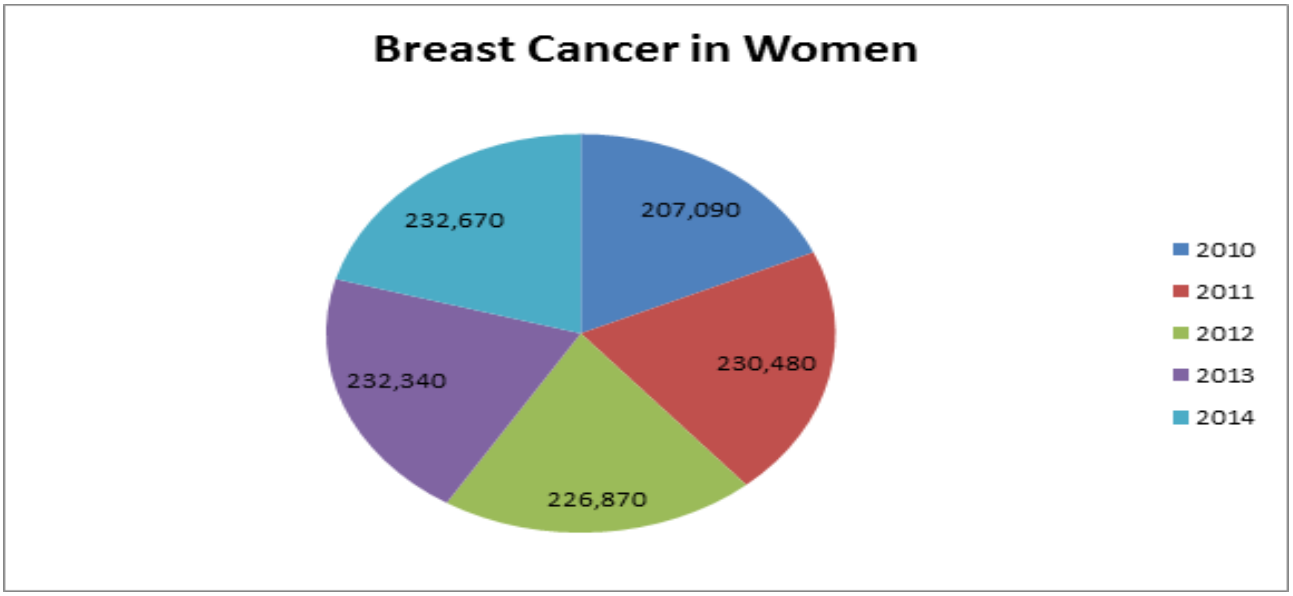


Figure 5.4 Breast Cancer Cases (2010-2014)

The following data sheet is a summary of all the data that has been processed:

Table 5.6:Summary of all processed Data (Male&Female)

Cancer Type	MALE NEW TOTAL	FEMALE NEW TOTAL	MALE DEATH TOTAL	FEMALE DEATH TOTAL
Breast	10,900	1,129,450	2,090	198,490
Lung & bronchus	582,360	539,850	433,760	359,560
Colon†	246,870	259,850	130,870	122,710
Uterine corpus	0	239,260	0	40,860
Thyroid	65,560	206,790	3,910	5,040
Melanoma-Skin	212,080	155,320	30,230	15,630
Non-Hodgkin Lymphoma	154,230	130,810	42,850	35,610
Kidney & renal pelvis	192,310	120,690	42,810	24,460
Pancreas	111,780	110,950	96,630	93,270
Ovary	0	110,370	0	73,110
Leukemia	134,820	100,970	66,600	48,370
Urinary Bladder	271,380	89,170	53,580	21,760
Rectum	116,000	84,170	0	0
Uterine cervix	0	61,780	0	20,770
Brain & other nervous system	62,460	51,320	38,600	29,750
Myeloma	60,700	48,100	29,730	24,040
Stomach	65,830	41,830	32,260	21,170
Liver & intrahepatic bile duct	105,380	37,480	70,720	33,000
Gallbladder & other biliary	22,620	27,160	6,580	10,100
Soft tissue (including heart)	30,680	25,530	11,180	9,690
Mouth	34,290	23,000	5,550	3,730
Vulva	0	22,280	0	4,830
Anus, anal canal, & anorectum	11,680	19,900	1,580	2,520

The analysis of problem #2 can be found in section 6- Analysis.

Problem #3 Which cancer is deadliest for male and female in the US. For this problem, we used the death cases data for the last 5 years.

Table 5.7:Deadliest Cancer in the US (Male & Female)

Cancer Type	2010 Male Death	2011 Male Death	2012 Male Death	2013 Male Death	2014 Male Death	2010 Female Death	2011 Female Death	2012 Female Death	2013 Female Death	2014 Female Death
Tongue	1,300	1,320	1360	1,380	1,450	690	710	690	690	700
Mouth	1,140	1,130	1,070	1,080	1,130	690	660	720	720	940
Pharynx	1,730	1,740	1730	1,790	1,900	680	690	600	610	640
Other oral cavity	1,260	1,270	1280	1,260	1,250	390	380	400	380	380
Esophagus	11,650	11,910	12040	12,220	12,450	2,850	2,800	3030	2,990	3,000
Stomach	6,350	6,260	6,190	6,740	6,720	4,220	4,080	4,350	4,250	4,270
Small intestine	610	610	610	610	640	490	490	540	560	570
Colon†	26,580	25,250	26,470	26,300	26,270	24,790	24,130	25,220	24,530	24,040
Rectum										
Anus, anal canal, & anorectum	280	300	300	330	370	440	470	480	550	580
Liver & intrahepatic bile duct	12,720	13,260	13,980	14,890	15,870	6,190	6,330	6,570	6,780	7,130
Gallbladder & other biliary	1,240	1,230	1,240	1,260	1,610	2,080	2,070	1,960	1,970	2,020
Pancreas	18,770	19,360	18850	19,480	20,170	18,030	18,300	18540	18,980	19,420
Other digestive organs	810	840	880	870	870	1,480	1,560	1,260	1,260	1,260
Larynx	2,870	2,840	2880	2,860	2,870	730	720	770	770	740
Lung & bronchus	86,220	85,600	87750	87,260	86,930	71,080	71,340	72,590	72,220	72,330
Other respiratory organs	460	450	480	480	480	310	300	300	300	310
Bones & joints	830	850	790	810	830	630	640	620	630	630
Soft tissue (including heart)	2,020	2,060	2050	2,500	2,550	1,900	1,860	1,850	1,890	2,190
Melanoma-Skin	5,670	5,750	6,060	6,280	6,470	3,030	3,040	3,120	3,200	3,240
Breast	390	450	410	410	430	39,840	39,520	39,510	39,620	40,000
Uterine cervix						4,210	4,290	4,220	4,030	4,020

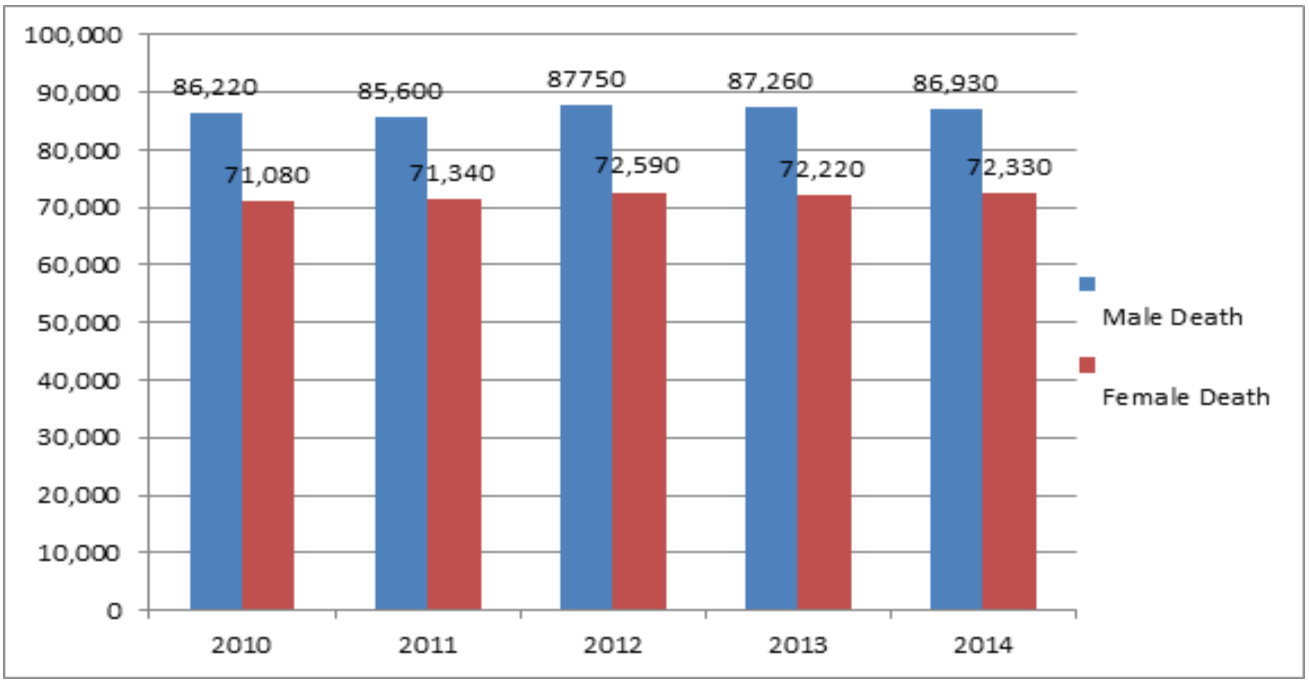


Figure 5.5 Death due to Lung & Bronchus Cancer -Male&Female (2010-2014)

Outcome of the summarized data sheets:

Table 5.8:Summary of Cancer Data based on Death Cases in the US (Male & Female)

Cancer Type	MALE NEW TOTAL	FEMALE NEW TOTAL	MALE DEATH TOTAL	FEMALE DEATH TOTAL
Breast	10,900	1,129,450	2,090	198,490
Lung & bronchus	582,360	539,850	433,760	359,560
Colon†	246,870	259,850	130,870	122,710
Uterine corpus	0	239,260	0	40,860
Thyroid	65,560	206,790	3,910	5,040
Melanoma-Skin	212,080	155,320	30,230	15,630
Non-Hodgkin Lymphoma	154,230	130,810	42,850	35,610
Kidney & renal pelvis	192,310	120,690	42,810	24,460
Pancreas	111,780	110,950	96,630	93,270
Ovary	0	110,370	0	73,110
Leukemia	134,820	100,970	66,600	48,370
Urinary Bladder	271,380	89,170	53,580	21,760
Rectum	116,000	84,170	0	0
Uterine cervix	0	61,780	0	20,770
Brain & other nervous system	62,460	51,320	38,600	29,750
Myeloma	60,700	48,100	29,730	24,040
Stomach	65,830	41,830	32,260	21,170
Liver & intrahepatic bile duct	105,380	37,480	70,720	33,000
Gallbladder & other biliary	22,620	27,160	6,580	10,100
Soft tissue (including heart)	30,680	25,530	11,180	9,690
Mouth	34,290	23,000	5,550	3,730
Vulva	0	22,280	0	4,830
Anus, anal canal, & anorectum	11,680	19,900	1,580	2,520

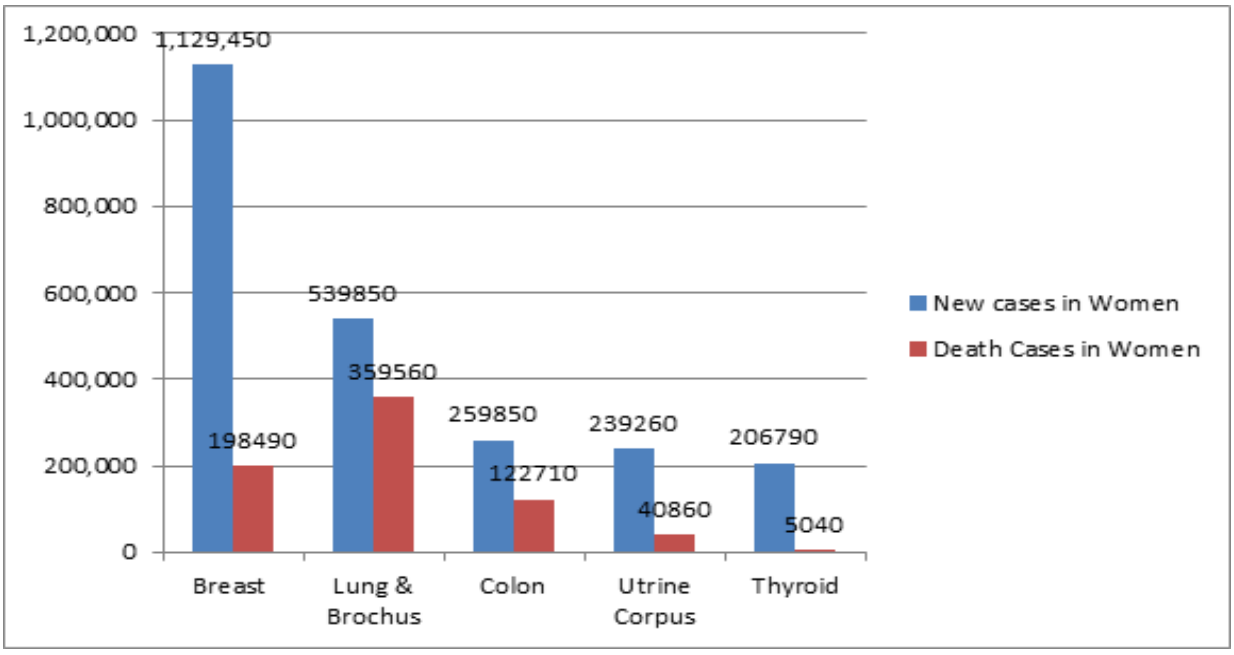


Figure 5.6 Death Cases and New Cases due to Lung & Bronchus Cancer -Female (2010-2014)

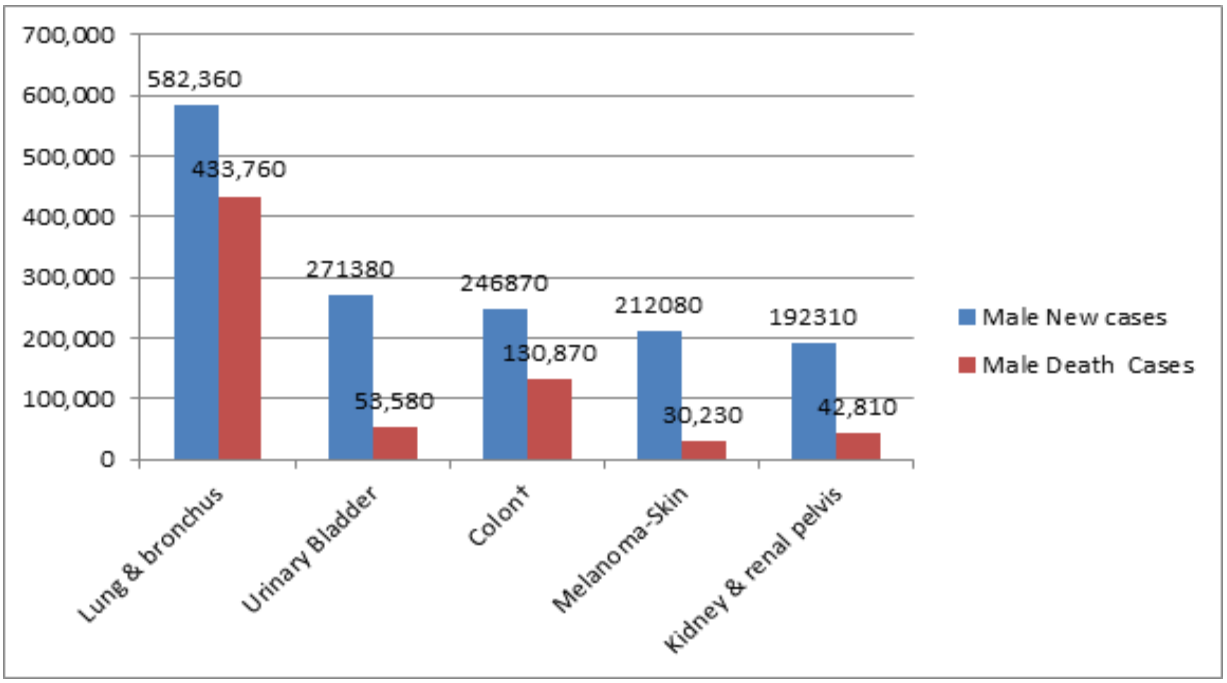


Figure 5.7 Death Cases and New Cases due to Lung & Bronchus Cancer -Male (2010-2014)

The analysis of problem #3 can be found in section 6- Analysis.

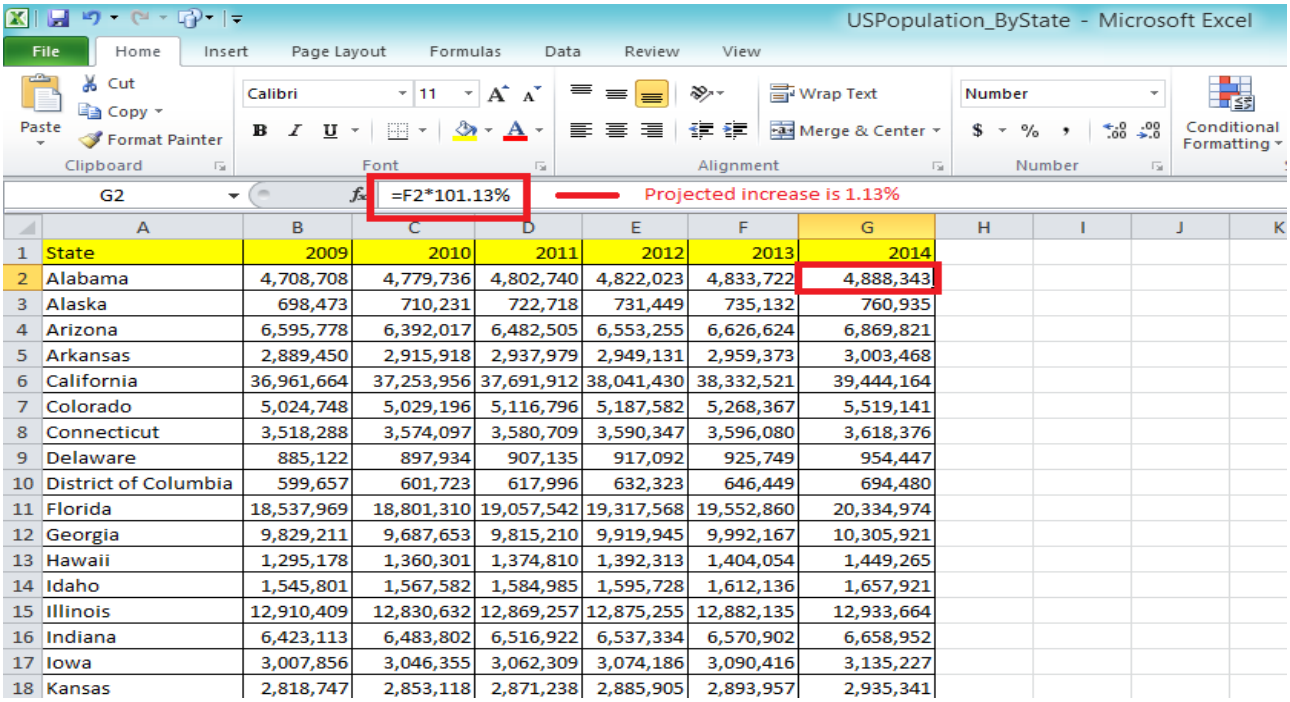
6. Analysis

Data Analysis has been conducted using MS Excel tool , it has been utilised to transform data using mathematical calculations such as the ones to obtain population and other pivot table analysis.

6.1 Problem #1

For our first problem, in order to derive maximum risk zone for each state, we need to consider the state by state population of the United States between 2010 to 2014. The website(census.gov) provides state wise population of the US between the years 2010 and 2013 however, for the happening year (2014) we do not have the fact sheets. Hence we use mathematically predict the population for all the states individually for 2014. This had to be done in order to complete the cancer data analysis for the year 2014 along with the other year's data. The projection calculation has been referenced from wikipedia.

Link - http://en.wikipedia.org/wiki/List_of_U.S._states_by_population_growth_rate.



	A	B	C	D	E	F	G	H	I	J	K
1	State	2009	2010	2011	2012	2013	2014				
2	Alabama	4,708,708	4,779,736	4,802,740	4,822,023	4,833,722	4,888,343				
3	Alaska	698,473	710,231	722,718	731,449	735,132	760,935				
4	Arizona	6,595,778	6,392,017	6,482,505	6,553,255	6,626,624	6,869,821				
5	Arkansas	2,889,450	2,915,918	2,937,979	2,949,131	2,959,373	3,003,468				
6	California	36,961,664	37,253,956	37,691,912	38,041,430	38,332,521	39,444,164				
7	Colorado	5,024,748	5,029,196	5,116,796	5,187,582	5,268,367	5,519,141				
8	Connecticut	3,518,288	3,574,097	3,580,709	3,590,347	3,596,080	3,618,376				
9	Delaware	885,122	897,934	907,135	917,092	925,749	954,447				
10	District of Columbia	599,657	601,723	617,996	632,323	646,449	694,480				
11	Florida	18,537,969	18,801,310	19,057,542	19,317,568	19,552,860	20,334,974				
12	Georgia	9,829,211	9,687,653	9,815,210	9,919,945	9,992,167	10,305,921				
13	Hawaii	1,295,178	1,360,301	1,374,810	1,392,313	1,404,054	1,449,265				
14	Idaho	1,545,801	1,567,582	1,584,985	1,595,728	1,612,136	1,657,921				
15	Illinois	12,910,409	12,830,632	12,869,257	12,875,255	12,882,135	12,933,664				
16	Indiana	6,423,113	6,483,802	6,516,922	6,537,334	6,570,902	6,658,952				
17	Iowa	3,007,856	3,046,355	3,062,309	3,074,186	3,090,416	3,135,227				
18	Kansas	2,818,747	2,853,118	2,871,238	2,885,905	2,893,957	2,935,341				

Figure 6.1 State Population 2014 Computation

A. For our first problem, we used comparative study for the last 5 years of "new cases", using the ratio of "All Sites" against "State Population" (partial data is illustrated).

Table 6.1: 5-year comparative study of new cancer cases

State	Zone	2010	2011	2012	2013	2014	AVERAGE
Alabama	Central	0.4946%	0.5316%	0.5483%	0.5602%	0.5476%	0.5365%
Alaska	Alaskan	0.4027%	0.4276%	0.4976%	0.4475%	0.4928%	0.4536%
Arizona	Mountain	0.4659%	0.4867%	0.4882%	0.5132%	0.4779%	0.4864%
Arkansas	Central	0.5254%	0.5470%	0.5466%	0.5518%	0.5500%	0.5442%
California	Pacific	0.4223%	0.4337%	0.4359%	0.4470%	0.4354%	0.4348%
Colorado	Mountain	0.4243%	0.4376%	0.4399%	0.4444%	0.4314%	0.4355%
Connecticut	Eastern	0.5806%	0.5988%	0.5997%	0.5890%	0.6099%	0.5956%
Delaware	Eastern	0.5446%	0.5655%	0.5823%	0.5801%	0.5574%	0.5660%
District of Columbia	Eastern	0.4587%	0.4579%	0.4713%	0.4517%	0.4089%	0.4497%
Florida	Eastern	0.5691%	0.5950%	0.6087%	0.6051%	0.5634%	0.5883%
Georgia	Eastern	0.4179%	0.4542%	0.4852%	0.4932%	0.4598%	0.4620%
Hawaii	Hawaiian	0.4903%	0.4881%	0.4747%	0.4736%	0.4582%	0.4770%
Idaho	Mountain	0.4606%	0.4745%	0.4838%	0.4758%	0.4819%	0.4753%
Illinois	Central	0.4979%	0.5098%	0.5107%	0.5130%	0.5168%	0.5097%
Indiana	Eastern	0.5093%	0.5225%	0.5363%	0.5410%	0.5340%	0.5286%

Taking the average of the the 5 years of data, Maine has been most affected by cancer from the data collected from “new cancer cases”. Hence **Maine is at high risk zone for cancer affected states**.

Table 6.2: High risk zone for cancer- Maine

State	Zone	2010	2011	2012	2013	2014	AVERAGE
Maine	Eastern	0.6512%	0.6641%	0.6764%	0.6919%	0.6901%	0.6747%
Vermont	Eastern	0.5945%	0.6306%	0.6486%	0.6703%	0.6582%	0.6404%
New Hampshire	Eastern	0.5933%	0.6228%	0.6322%	0.6400%	0.6351%	0.6247%
Pennsylvania	Eastern	0.5925%	0.6123%	0.6138%	0.6228%	0.6222%	0.6127%
West Virginia	Eastern	0.5726%	0.5972%	0.6257%	0.6175%	0.6305%	0.6087%
Connecticut	Eastern	0.5806%	0.5988%	0.5997%	0.5890%	0.6099%	0.5956%
Rhode Island	Eastern	0.5672%	0.5793%	0.6008%	0.5972%	0.6052%	0.5899%
Florida	Eastern	0.5691%	0.5950%	0.6087%	0.6051%	0.5634%	0.5883%
Michigan	Eastern	0.5632%	0.5772%	0.5847%	0.5817%	0.5916%	0.5797%
Kentucky	Eastern	0.5586%	0.5724%	0.5744%	0.5711%	0.5788%	0.5711%

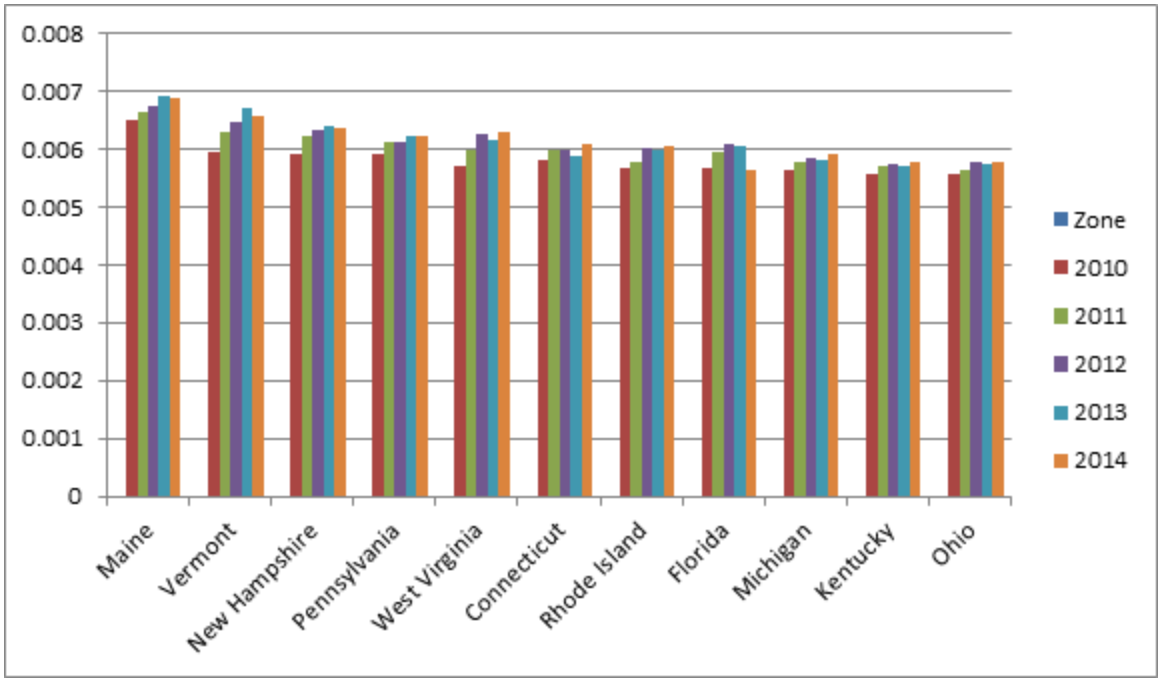


Figure 6.2:High Risk zone for cancer affected States

The reason for Maine state being the most affected by cancer is out of the scope of our study, as that deals with data related to treatment of patients , availability of resources, exposure to cancer causing elements, awareness to tests and treatments to quotes a few reasons.

Initial Assumption: California is the State with the maximum risk zone for cancer , this assumption was made because of the number of "new cases" data.

Factual Deduction: Upon processing and comparison California is ranked 49 amongst maximum risk zone for cancer.

Table 6.3: Assumption Tested and proven Wrong

Louisiana	Central	0.4621%	0.4979%	0.5102%	0.5390%	0.5149%	0.5048%
Maryland	Eastern	0.4798%	0.4957%	0.5268%	0.5175%	0.5039%	0.5047%
Nebraska	Central	0.5054%	0.5118%	0.4867%	0.4849%	0.4996%	0.4976%
North Dakota	Central	0.4906%	0.5205%	0.5017%	0.4852%	0.4794%	0.4955%
Kansas	Central	0.4749%	0.4900%	0.4882%	0.4966%	0.4984%	0.4896%
Arizona	Mountain	0.4659%	0.4867%	0.4882%	0.5132%	0.4779%	0.4864%
Nevada	Pacific	0.4529%	0.4700%	0.4995%	0.4957%	0.5013%	0.4839%
Virginia	Eastern	0.4551%	0.4782%	0.5055%	0.4948%	0.4804%	0.4828%
Hawaii	Hawaiian	0.4903%	0.4881%	0.4747%	0.4736%	0.4582%	0.4770%
Idaho	Mountain	0.4606%	0.4745%	0.4838%	0.4758%	0.4819%	0.4753%
New Mexico	Mountain	0.4473%	0.4625%	0.4622%	0.4839%	0.4835%	0.4679%
Wyoming	Mountain	0.4507%	0.4717%	0.4597%	0.4634%	0.4798%	0.4651%
Georgia	Eastern	0.4179%	0.4542%	0.4852%	0.4932%	0.4598%	0.4620%
Alaska	Alaskan	0.4027%	0.4276%	0.4976%	0.4475%	0.4928%	0.4536%
District of Columbia	Eastern	0.4587%	0.4579%	0.4713%	0.4517%	0.4089%	0.4497%
Colorado	Mountain	0.4243%	0.4376%	0.4399%	0.4444%	0.4314%	0.4355%
California	Pacific	0.4223%	0.4337%	0.4359%	0.4470%	0.4354%	0.4348%
Texas	Central	0.4021%	0.4090%	0.4239%	0.4243%	0.4160%	0.4151%
Utah	Mountain	0.3607%	0.3738%	0.3719%	0.3726%	0.3541%	0.3666%

Considering data for different zones, We conclude that most new cases are reported on the Eastern Zone. This conclusion was drawn based on the average of AVERAGE done on the above data for different zones.

Pivot Table 6.1:Average of AVERAGE

Row Labels	Average of AVERAGE
Alaskan	0.4536%
Central	0.5148%
Central/Mountain	0.5281%
Eastern	0.5627%
Hawaiian	0.4770%
Mountain	0.4648%
Pacific	0.4978%

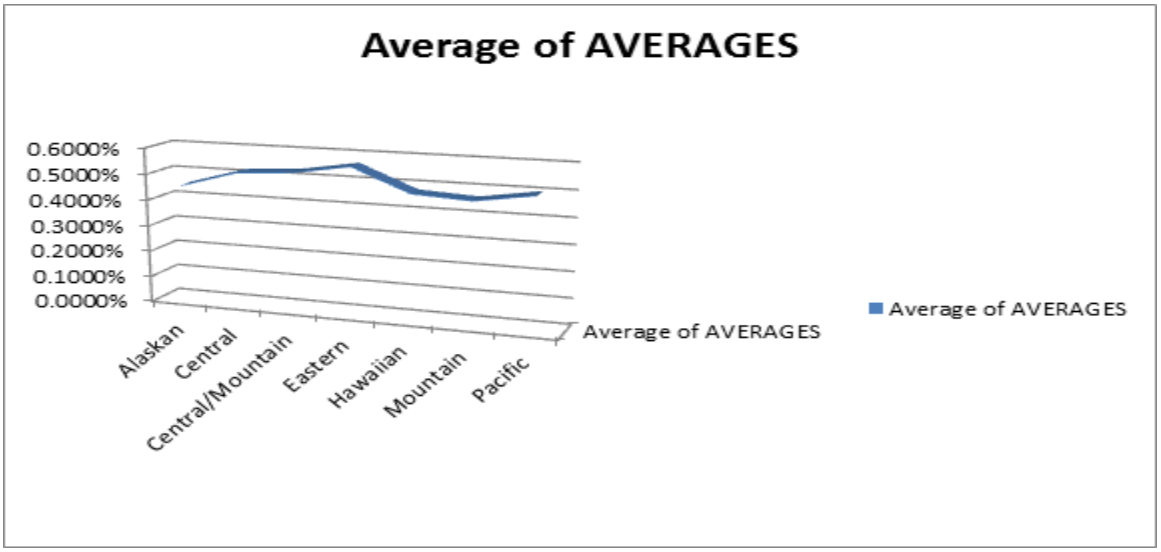


Figure 6.3: Average of AVERAGE for different Zones

B. The second part of first problem deals with the comparative study. A comparative study on the 5 years of data on death cases is done by computing the ratio of "All Sites" against "State Population" as shown below (partial data is illustrated)

Table 6.4: Comparative study of 5 year data on Death cases in the US

State	Zone	2010	2011	2012	2013	2014	AVERAGE
Alabama	Central	0.2124%	0.2126%	0.2134%	0.2134%	0.2150%	0.2133%
Alaska	Alaskan	0.1239%	0.1259%	0.1271%	0.1288%	0.1301%	0.1272%
Arizona	Mountain	0.1663%	0.1669%	0.1692%	0.1632%	0.1659%	0.1663%
Arkansas	Central	0.2215%	0.2199%	0.2228%	0.2214%	0.2241%	0.2219%
California	Pacific	0.1495%	0.1487%	0.1488%	0.1452%	0.1469%	0.1478%
Colorado	Mountain	0.1368%	0.1364%	0.1386%	0.1332%	0.1355%	0.1361%
Connecticut	Eastern	0.1917%	0.1899%	0.1933%	0.1904%	0.1901%	0.1911%
Delaware	Eastern	0.2116%	0.2128%	0.2104%	0.2033%	0.2074%	0.2091%
District of Columbia	Eastern	0.1595%	0.1489%	0.1597%	0.1483%	0.1454%	0.1524%
Florida	Eastern	0.2174%	0.2150%	0.2183%	0.2084%	0.2102%	0.2139%
Georgia	Eastern	0.1607%	0.1616%	0.1592%	0.1553%	0.1584%	0.1590%
Hawaii	Hawaiian	0.1713%	0.1724%	0.1709%	0.1656%	0.1691%	0.1699%
Idaho	Mountain	0.1614%	0.1621%	0.1654%	0.1604%	0.1647%	0.1628%
Illinois	Central	0.1821%	0.1798%	0.1862%	0.1856%	0.1857%	0.1839%
Indiana	Eastern	0.1990%	0.1989%	0.2025%	0.1990%	0.2008%	0.2000%

West Virginia is the state most affected by deaths that is caused by cancer. This conclusion has been made by taking the average of the 5 years of "death cases" data and comparing that data.

Table 6.5: Most affected State in the US

State	Zone	2010	2011	2012	2013	2014	AVERAGE
West Virginia	Eastern	0.2520%	0.2522%	0.2479%	0.2511%	0.2522%	0.2511%
Maine	Eastern	0.2386%	0.2394%	0.2430%	0.2412%	0.2457%	0.2416%
Kentucky	Eastern	0.2228%	0.2231%	0.2258%	0.2239%	0.2275%	0.2246%
Pennsylvania	Eastern	0.2259%	0.2241%	0.2256%	0.2233%	0.2232%	0.2244%
Arkansas	Central	0.2215%	0.2199%	0.2228%	0.2214%	0.2241%	0.2219%
Ohio	Eastern	0.2165%	0.2157%	0.2168%	0.2165%	0.2177%	0.2166%
Tennessee	Central	0.2143%	0.2154%	0.2150%	0.2118%	0.2148%	0.2142%
Florida	Eastern	0.2174%	0.2150%	0.2183%	0.2084%	0.2102%	0.2139%
Alabama	Central	0.2124%	0.2126%	0.2134%	0.2134%	0.2150%	0.2133%
Missouri	Central	0.2107%	0.2113%	0.2111%	0.2087%	0.2110%	0.2106%

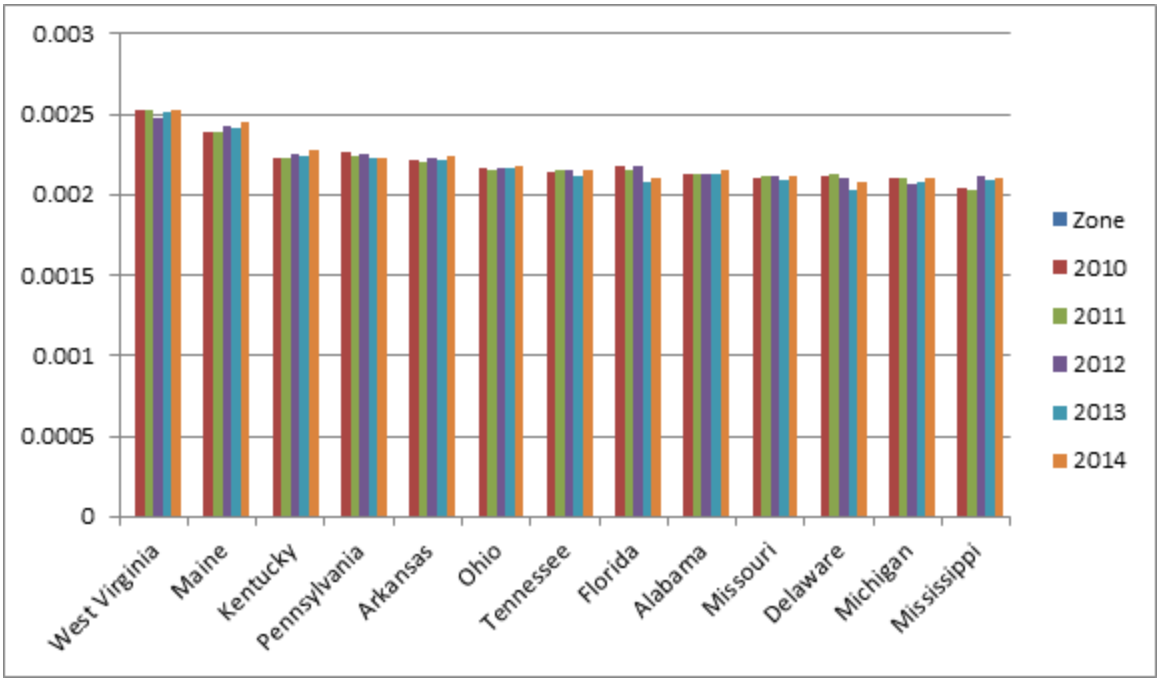


Figure 6.4: Death cases in the US due to cancer

*Initial Assumption :*California is the State with the maximum risk zone for deaths caused by cancer , this assumption was made because of the number of "death cases" data.

Factual Deduction: Upon processing and comparison California is ranked 48 amongst maximum risk zone for cancer caused deaths in the state.

Table 6.6: Assumption Tested and Proven Wrong

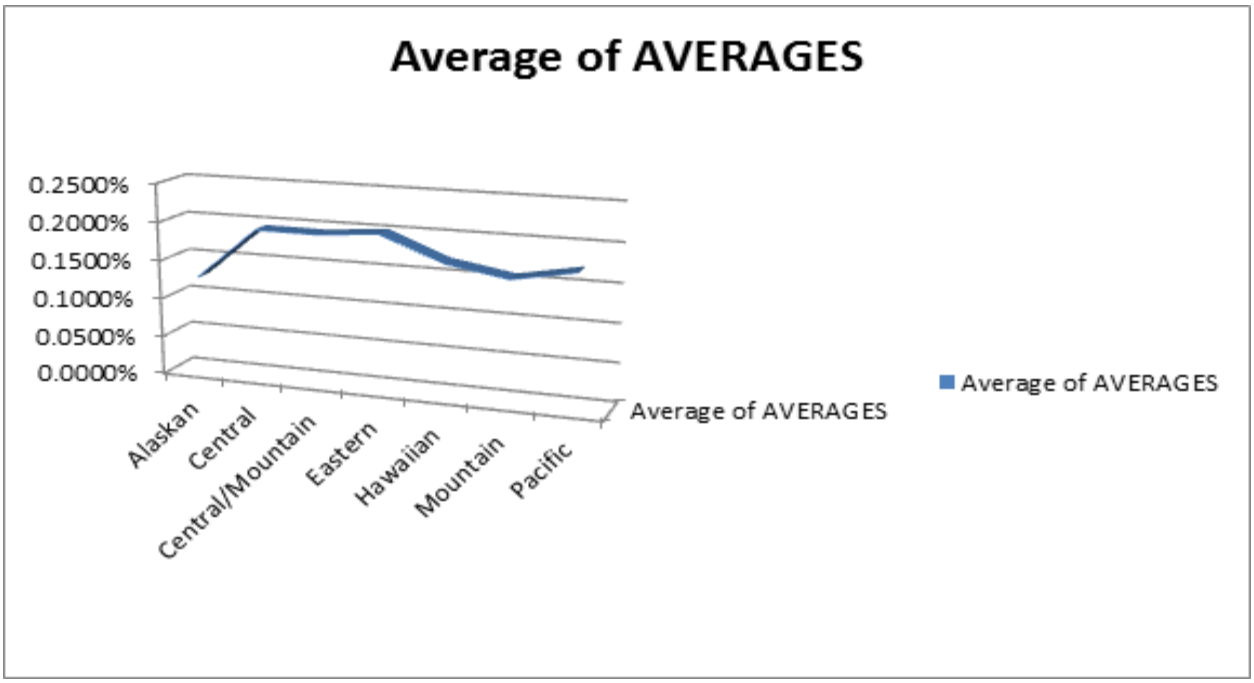
Nebraska	Central	0.1916%	0.1905%	0.1859%	0.1799%	0.1820%	0.1860%
New Jersey	Eastern	0.1879%	0.1856%	0.1878%	0.1822%	0.1815%	0.1850%
Illinois	Central	0.1821%	0.1798%	0.1862%	0.1856%	0.1857%	0.1839%
North Dakota	Central	0.1903%	0.1872%	0.1858%	0.1645%	0.1632%	0.1782%
Virginia	Eastern	0.1779%	0.1771%	0.1785%	0.1726%	0.1730%	0.1758%
Maryland	Eastern	0.1775%	0.1757%	0.1774%	0.1721%	0.1725%	0.1750%
New York	Eastern	0.1782%	0.1765%	0.1744%	0.1718%	0.1728%	0.1748%
Minnesota	Central	0.1735%	0.1729%	0.1764%	0.1735%	0.1760%	0.1744%
Washington	Pacific	0.1731%	0.1719%	0.1765%	0.1714%	0.1736%	0.1733%
Hawaii	Hawaiian	0.1713%	0.1724%	0.1709%	0.1656%	0.1691%	0.1699%
Nevada	Pacific	0.1718%	0.1741%	0.1664%	0.1651%	0.1662%	0.1687%
Wyoming	Mountain	0.1774%	0.1795%	0.1631%	0.1577%	0.1644%	0.1684%
New Mexico	Mountain	0.1651%	0.1662%	0.1693%	0.1676%	0.1705%	0.1677%
Arizona	Mountain	0.1663%	0.1669%	0.1692%	0.1632%	0.1659%	0.1663%
Idaho	Mountain	0.1614%	0.1621%	0.1654%	0.1604%	0.1647%	0.1628%
Georgia	Eastern	0.1607%	0.1616%	0.1592%	0.1553%	0.1584%	0.1590%
District of Columbia	Eastern	0.1595%	0.1489%	0.1597%	0.1483%	0.1454%	0.1524%
California	Pacific	0.1495%	0.1487%	0.1488%	0.1452%	0.1469%	0.1478%
Texas	Central	0.1453%	0.1432%	0.1413%	0.1337%	0.1360%	0.1399%
Colorado	Mountain	0.1368%	0.1364%	0.1386%	0.1332%	0.1355%	0.1361%
Alaska	Alaskan	0.1239%	0.1259%	0.1271%	0.1288%	0.1301%	0.1272%
Utah	Mountain	0.1020%	0.1022%	0.0974%	0.0916%	0.0943%	0.0975%

Considering data for different zones, We conclude that most deaths are reported on the Eastern Zone. This conclusion was drawn based on the average of AVERAGE done on the above data for different zones.

Row Labels	Average of AVERAGE
Alaskan	0.1272%
Central	0.1943%
Central/Mountain	0.1939%
Eastern	0.2002%
Hawaiian	0.1699%
Mountain	0.1565%
Pacific	0.1715%

Pivot Table 6.2: Average of AVERAGES

Figure 6.5: Average of AVERAGES



6.2 Problem #2

To detect the most prevalent type of cancer in the US we use "new cases" data of the 5 year data .(Also , shown on Section 5 analysis).

Table 6.7: 5 year analysis of New cases of cancer in Men in the US

Cancer Type	MALE NEW TOTAL	FEMALE NEW TOTAL	MALE DEATH TOTAL	FEMALE DEATH TOTAL
Prostate	1,171,950	0	153,140	0
Lung & bronchus	582,360	539,850	433,760	359,560
Urinary Bladder	271,380	89,170	53,580	21,760
Colon†	246,870	259,850	130,870	122,710
Melanoma-Skin	212,080	155,320	30,230	15,630
Kidney & renal pelvis	192,310	120,690	42,810	24,460
Non-Hodgkin Lymphoma	154,230	130,810	42,850	35,610
Leukemia	134,820	100,970	66,600	48,370
Rectum	116,000	84,170	0	0
Pancreas	111,780	110,950	96,630	93,270
Liver & intrahepatic bile duct	105,380	37,480	70,720	33,000
Esophagus	69,630	17,610	60,270	14,670
Stomach	65,830	41,830	32,260	21,170
Thyroid	65,560	206,790	3,910	5,040
Brain & other nervous system	62,460	51,320	38,600	29,750
Myeloma	60,700	48,100	29,730	24,040
Pharynx	54,020	14,070	8,890	3,220
Larynx	49,790	12,920	14,320	3,730
Tongue	44,910	18,090	6,810	3,480
Testis	42,100	0	1,810	0
Mouth	34,290	23,000	5,550	3,730
Soft tissue (including heart)	30,680	25,530	11,180	9,690

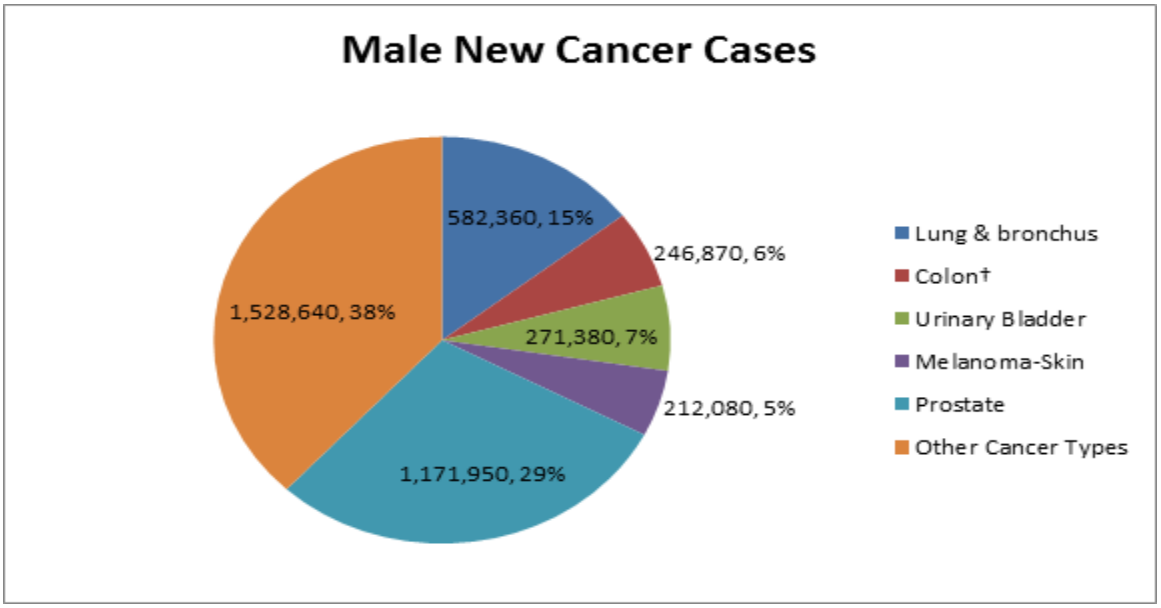


Figure 6.6: New cancer cases in Men in the US

Hypothesis Testing:

Initial Assumption : Prostate cancer will be the most dominant cancer in the coming years for men.

Factual Deduction: Upon processing the data in hand, the data is in agreement with the assumption that Prostate cancer will be the most dominant cancer in the coming years that would affect men.

Table 6.8: 5 year analysis of New cases of cancer in Women in the US

Cancer Type	MALE NEW	FEMALE NEW	MALE DEATH	FEMALE DEATH
	TOTAL	TOTAL	TOTAL	TOTAL
Breast	10,900	1,129,450	2,090	198,490
Lung & bronchus	582,360	539,850	433,760	359,560
Colon†	246,870	259,850	130,870	122,710
Uterine corpus	0	239,260	0	40,860
Thyroid	65,560	206,790	3,910	5,040
Melanoma-Skin	212,080	155,320	30,230	15,630
Non-Hodgkin Lymphoma	154,230	130,810	42,850	35,610
Kidney & renal pelvis	192,310	120,690	42,810	24,460
Pancreas	111,780	110,950	96,630	93,270
Ovary	0	110,370	0	73,110
Leukemia	134,820	100,970	66,600	48,370
Urinary Bladder	271,380	89,170	53,580	21,760
Rectum	116,000	84,170	0	0
Uterine cervix	0	61,780	0	20,770
Brain & other nervous system	62,460	51,320	38,600	29,750
Myeloma	60,700	48,100	29,730	24,040
Stomach	65,830	41,830	32,260	21,170
Liver & intrahepatic bile duct	105,380	37,480	70,720	33,000
Gallbladder & other biliary	22,620	27,160	6,580	10,100
Soft tissue (including heart)	30,680	25,530	11,180	9,690
Mouth	34,290	23,000	5,550	3,730
Vulva	0	22,280	0	4,830

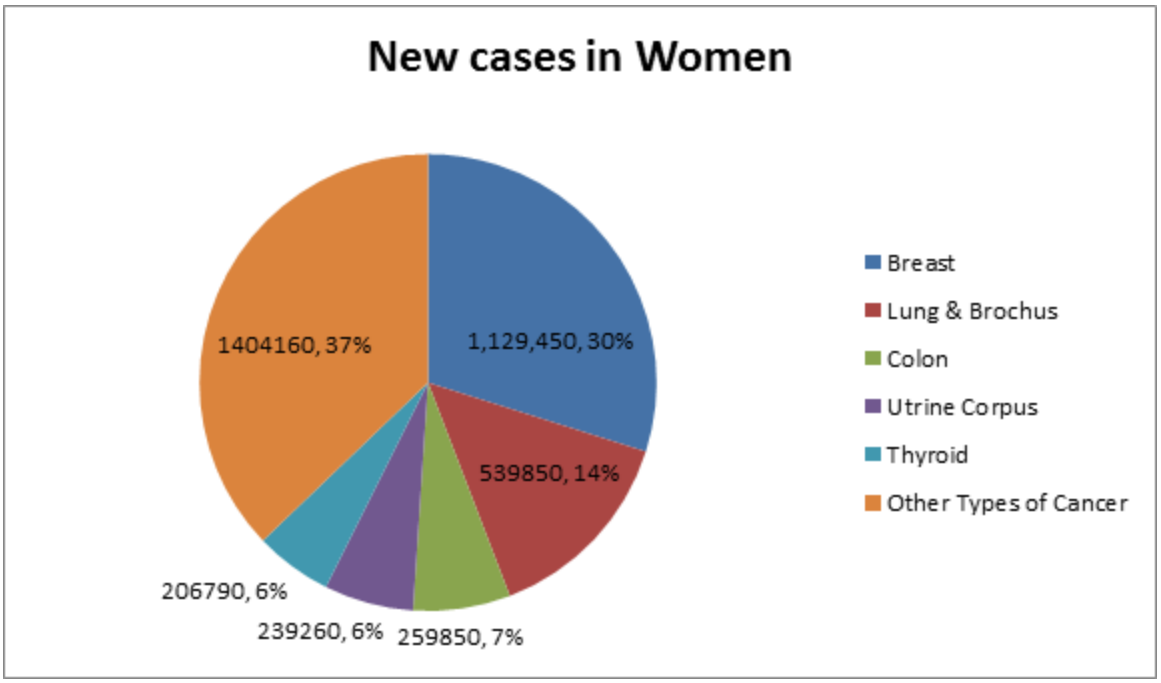


Figure 6.6: New cancer cases in Women in the US

Hypothesis Testing:

Initial Assumption : Breast cancer will be the most dominant cancer in the coming years for women.

Factual Deduction: Upon processing the data in hand, the data is in agreement with the assumption that Breast cancer will be the most dominant cancer in the coming years that would affect women.

Gender and Age based analysis of "new cases" in the US :

For males:

Table 6.9: Gender and age based analysis of New cases in Men in US

New cancer cases Year	New cancer cases All Sites	All sites men	New cancer cases % men	> 45 MEN	% >45 MEN	< 45 MEN	% < 45 MEN	>65 MEN	% > 65 MEN	< 65 MEN	%<65 MEN
2010	3,193,340	1,579,240	49.45	52,190	3.30	737,430	46.70	334,400	21.17	455,220	28.83
2011	3,277,820	1,644,600	50.17	52,870	3.21	769,430	46.79	354,660	21.57	467,640	28.43
2012	3,320,580	1,696,340	51.09	53,860	3.18	794,310	46.82	371,340	21.89	476,830	28.11
2013	3,331,080	1,709,580	51.32	53,290	3.12	801,500	46.88	376,950	22.05	477,840	27.95
2014	3,059,120	1,710,440	55.91	53,990	3.16	801,230	46.84	370,880	21.68	484,340	28.32

Age distribution of Men affected by Prostate cancer in the US is shown below :

Table 6.10: Age distribution of Men affected by Prostate Cancer in US

Prevalent new cancer in a given year	Total New Cancer cases in US	Men of all ages	Men under 45	% Men under 45	Men above 45	% Men above 45	Men under 65	% Men under 65	Men above 65	% Men above 65
Lungs (Men) 2010	445,040	233,500	2,120	0.91	114,630	49	38,950	16.68	77,800	33.32
Lungs (Men) 2011	442,260	230,120	1,910	0.83	113,150	49	38,290	16.64	76,770	33.36
Lungs (Men) 2012	452,320	232,940	1,740	0.75	114,730	49	38,760	16.64	77,710	33.36
Lungs (Men) 2013	456,380	236,160	1,590	0.67	116,490	49	38,390	16.26	79,690	33.74
Lungs (Men) 2014	448,420	232,000	1,690	0.73	114,310	49	38,190	16.46	77,810	33.54
Colon & Rectum (Men) 2010	285,140	144,180	3,480	2.41	68,610	48	28,370	19.68	43,720	30.32
Colon & Rectum (Men) 2011	282,420	143,700	3,640	2.53	68,210	47	29,420	20.47	42,430	29.53
Colon & Rectum (Men) 2012	193,020	52,940	880	1.66	25,590	48	8,510	16.07	17,960	33.93
Colon & Rectum (Men) 2013	215,840	77,560	3,810	4.91	70	0	31,610	40.76	42,070	54.24
Colon & Rectum (Men) 2014	273,660	143,660	3,680	2.56	68,150	47	30,160	20.99	41,670	29.01
Prostrate (Men) 2010	435,460	435,460	1,280	0.29	216,450	50	85,090	19.54	132,640	30.46
Prostrate (Men) 2011	481,780	481,780	1,470	0.31	239,420	50	97,920	20.32	142,970	29.68
Prostrate (Men) 2012	56,310	56,310	*		28,140	50	2,730	4.85	25,440	45.18
Prostrate (Men) 2013	477,180	477,180	1,450	0.30	237,140	50	103,990	21.79	134,600	28.21
Prostrate (Men) 2014	368,088	368,088	1,430	0.39	231,570	63	98	0.03	134,990	36.67

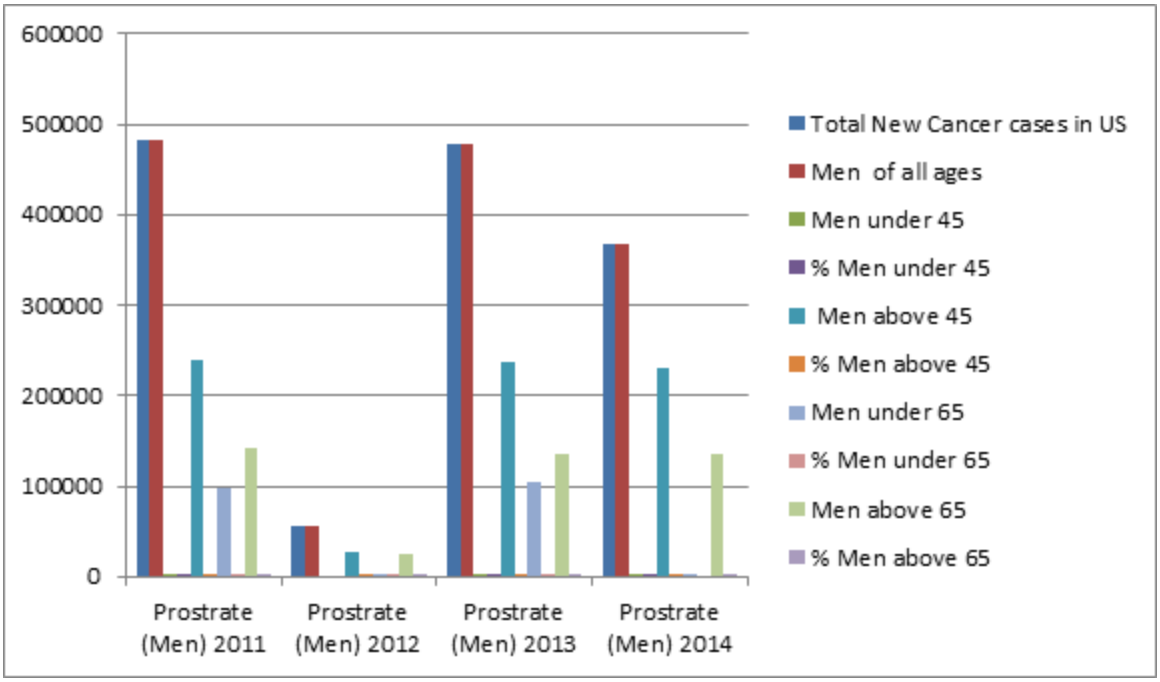


Figure 6.7: New cancer cases in Men in the US affected by Prostate Cancer

For woman:

Table 6.11: Age distribution of Women affected by Cancer in US

New cancer cases Year	New cancer cases All Sites	All sites women	Cancer deaths % female	> 45 WOMEN	% AGE >45 WOMEN	< 45 WOMEN	%AGE <45 WOMEN	>65 WOMEN	% AGE>65 WOMEN	<65 WOMEN	%AGE<65 WOMEN
2010	3,193,340	1,479,880	46.34	83760	5.66	656,180	44.34	352,510	23.82	387,430	26.18
2011	3,277,820	1,548,740	47.25	86,190	5.57	688,180	44.43	374,330	24.17	400,040	25.83
2012	3,320,580	1,581,480	47.63	85,870	5.43	704,870	44.57	384,260	24.30	406,480	25.70
2013	3,331,080	1,611,000	48.36	86,940	5.40	718,560	44.60	393,890	24.45	411,610	25.55
2014	3,059,120	1,620,640	52.98	87,920	5.43	722,400	44.57	390,910	24.12	419,410	25.88

Age distribution of women affected by Breast cancer in the US is shown below :

Table 6.12: Age distribution of Women affected by Breast Cancer in US

Prevalent new cancer in a given year	Total New Cancer cases in US	Women of all ages	Women under 45	% Women under 45	Women above 45	% Women above 45	Women under 65	% Women under 65	Women above 65	% Women above 65
Lungs (wOMen) 2010	445,040	211,540	2,450	1.16	103,320	48.84	34,250	16.19	71,520	33.81
Lungs (wOMen) 2011	442,260	212,140	2,200	1.04	103,870	48.96	34,300	16.17	71,770	33.83
Lungs (WOMen) 2012	452,320	219,380	2,060	0.94	107,630	49.06	35,010	15.96	74,680	34.04
Lungs (WOMen) 2013	456,380	220,220	1,980	0.90	108,130	49.10	35,100	15.94	75,010	34.06
Lungs (woMen) 2014	448,420	216,420	2,020	0.93	106,190	49.07	34,410	15.90	73,800	34.10
Colon & Rectum (woMen) 2010	285,140	140,960	3,280	2.33	67,200	47.67	22,790	16.17	47,690	33.83
Colon & Rectum (woMen) 2011	282,420	138,720	3,350	2.41	66,010	47.59	23,560	16.98	45,800	33.02
Colon & Rectum (woMen) 2012	193,020	140,080	3,460	2.47	66,580	47.53	24,190	17.27	45,850	32.73
Colon & Rectum (woMen) 2013	215,840	138,280	3,500	2.53	65,640	47.47	24,760	17.91	44,380	32.09
Colon & Rectum (woMen) 2014	273,660	130,000	3,260	2.51	61,740	47.49	22,820	17.55	42,180	32.45
Breast (woMen) 2010	137,060	137,060	24,520	17.89	6,170	4.50	18,360	13.40	88,010	64.21
Breast (woMen) 2011	460,960	460,960	26,390	5.73	204,090	44.27	132,870	28.82	97,610	21.18
Breast (woMen) 2012	453,740	453,740	24,610	5.42	202,260	44.58	130,000	28.65	96,870	21.35
Breast (woMen) 2013	464,680	464,680	24,430	5.26	207,910	44.74	132,480	28.51	99,860	21.49
Breast (woMen) 2014	465,340	465,340	25,500	5.48	207,170	44.52	133,310	28.65	99,360	21.35

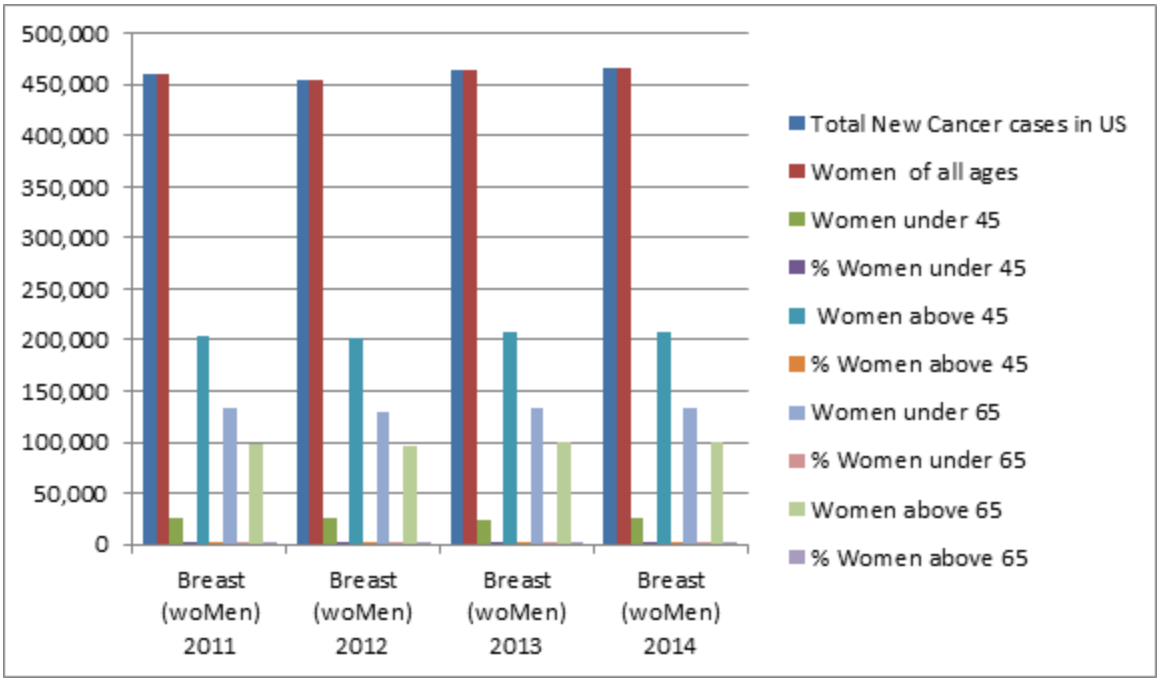


Figure 6.8: New cancer cases in Women in the US affected by Breast Cancer

6.3 Problem #3

To identify the types of cancer causing maximum deaths in the US in both men and woman, we use 5 years of "death cases" data (shown on Section- 5 Analysis).

Table 6.13: Analysis of Death cases in Men caused by Cancer with 5 years of data in US

Cancer Type	MALE NEW	FEMALE NEW	MALE DEATH	FEMALE DEATH
	TOTAL	TOTAL	TOTAL	TOTAL
Lung & bronchus	582,360	539,850	433,760	359,560
Prostate	1,171,950	0	153,140	0
Colon†	246,870	259,850	130,870	122,710
Pancreas	111,780	110,950	96,630	93,270
Liver & intrahepatic bile duct	105,380	37,480	70,720	33,000
Leukemia	134,820	100,970	66,600	48,370
Esophagus	69,630	17,610	60,270	14,670
Urinary Bladder	271,380	89,170	53,580	21,760
Non-Hodgkin Lymphoma	154,230	130,810	42,850	35,610
Kidney & renal pelvis	192,310	120,690	42,810	24,460
Brain & other nervous system	62,460	51,320	38,600	29,750
Stomach	65,830	41,830	32,260	21,170
Melanoma-Skin	212,080	155,320	30,230	15,630
Myeloma	60,700	48,100	29,730	24,040
Larynx	49,790	12,920	14,320	3,730
Soft tissue (including heart)	30,680	25,530	11,180	9,690
Pharynx	54,020	14,070	8,890	3,220
Tongue	44,910	18,090	6,810	3,480
Gallbladder & other biliary	22,620	27,160	6,580	10,100
Other oral cavity	8,290	3,340	6,320	1,930
Mouth	34,290	23,000	5,550	3,730
Other digestive organs	8,930	18,150	4,270	6,820

Figure 6.9: Top 5 most deadly Cancer Causing Deaths in Men in the US

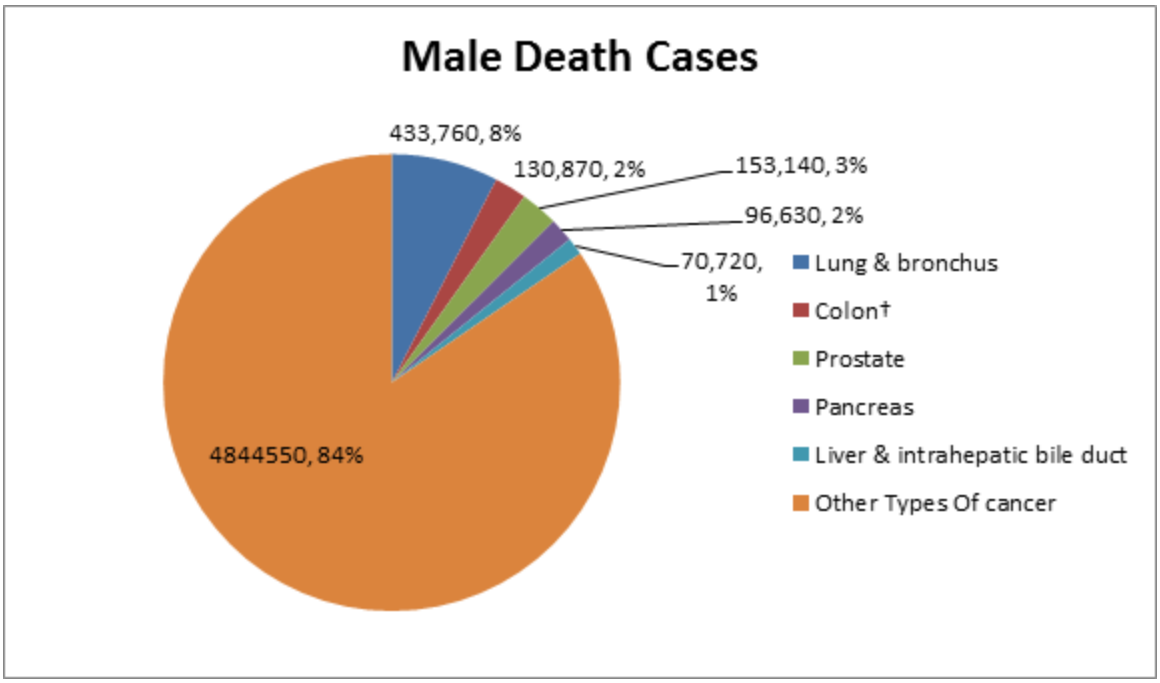
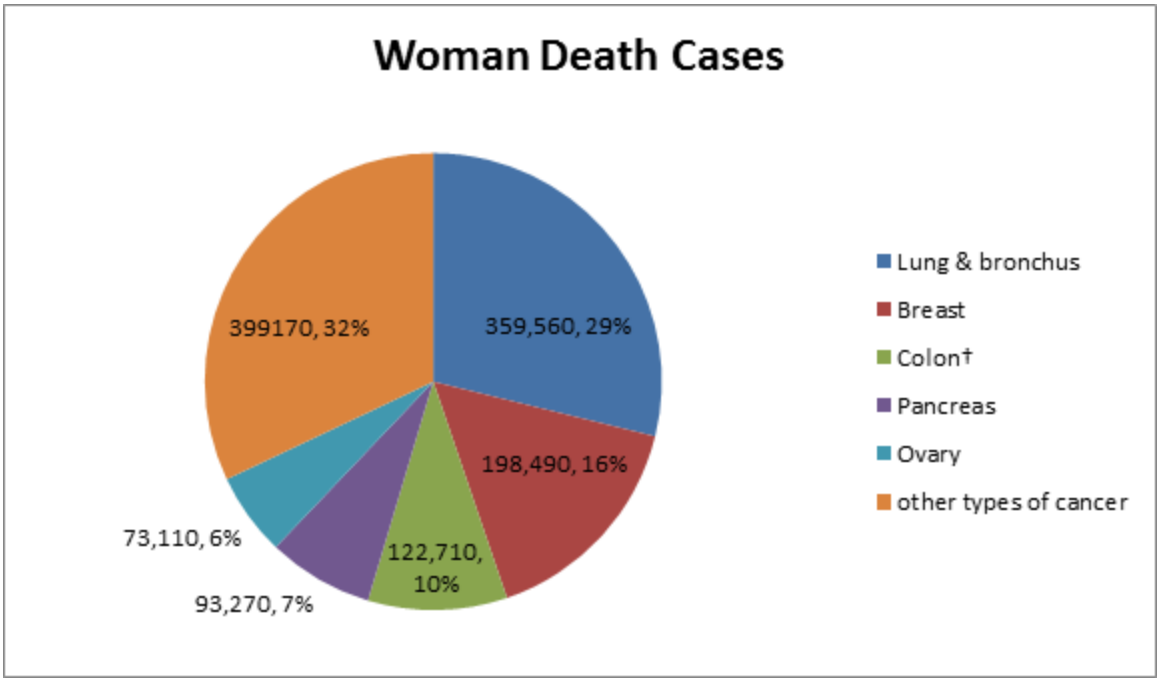


Figure 6.10: Top 5 most deadly Cancer Causing Deaths in Women in the US



Hypothesis Testing:

Initial Assumption : Lung and Bronchus cancer will be the most fatal cancer in the coming years for men and women in the US.

Factual Deduction: Upon processing the data in hand, the data is in agreement with the assumption that Lung and Bronchus cancer will be the most fatal cancer in the coming years for both men and women in the US.

Gender and Age based analysis of "death cases" in the US :

For males:

Table 6.14: Gender and age based analysis of Death cases in Men in US

Estimated cancer Deaths Year	Cancer deaths All Sites	All sites men	Cancer deaths % men	> 45 MEN	% >45 MEN	< 45 MEN	% < 45 MEN	>65 MEN	% > 65 MEN	< 65 MEN	%<65 MEN
2010	1,138,980	598,400	52.54	9,800	1.64	289,400	48.36	92,730	15.50	206,470	34.50
2011	1,143,900	600,860	52.53	9,840	1.64	290,590	48.36	93,110	15.50	207,320	34.50
2012	1,154,380	603,640	52.29	9,490	1.57	292,330	48.43	94,000	15.57	207,820	34.43
2013	1,151,700	604,840	52.52	9,370	1.55	297,550	49.19	95980	15.87	201940	33.39
2014	1,171,440	620,020	52.93	9490	1.53	300520	48.47	96920	15.63	213090	34.37

Age distribution of Men affected by Lung and Bronchus cancer in the US is shown below :

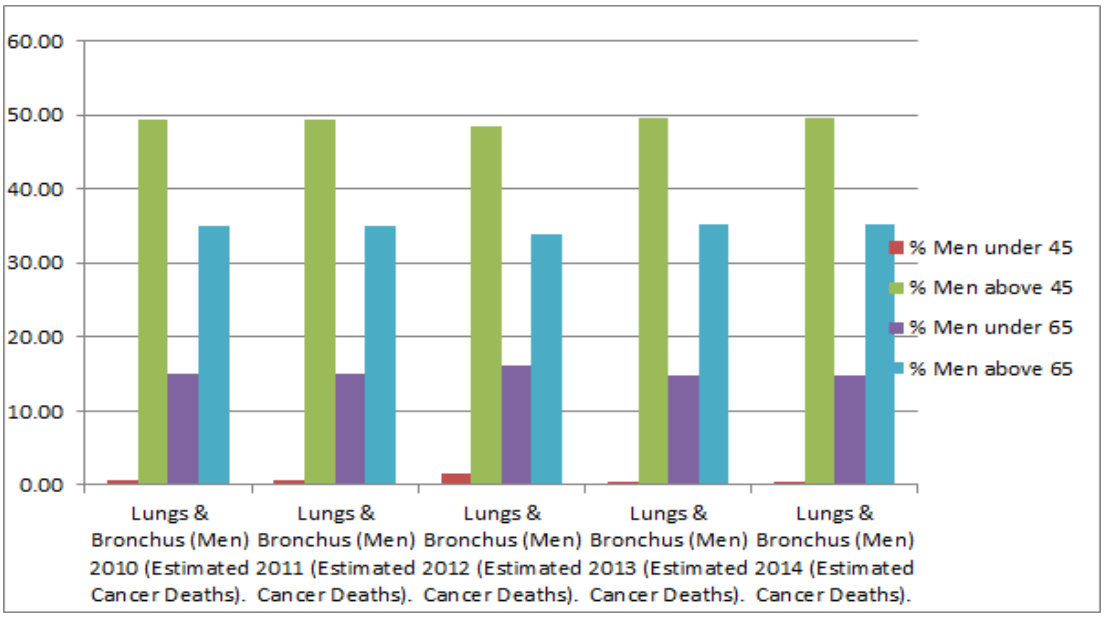
Table 6.15: Age distribution of Men affected by Lung & Bronchus Cancer in US

Prevalent Death cancer in a given year	Total Death Cancer cases in US	Men of all ages	Men under 45	% Men under 45	Men above 45	% Men above 45	Men under 65	% Men under 65	Men above 65	% Men above 65
Lungs & Bronchus (Men) 2010 (Estimated Cancer Deaths).	314,600	172,440	1,100	0.64	85,120	49.36	25,870	15.00	60,350	35.00
Lungs & Bronchus (Men) 2011 (Estimated Cancer Deaths).	313,880	171,200	1,100	0.64	84,500	49.36	25,690	15.01	59,910	34.99
Lungs & Bronchus (Men) 2012 (Estimated Cancer Deaths).	198,120	52,940	880	1.66	25,590	48.34	8,510	16.07	17,960	33.93
Lungs & Bronchus (Men) 2013 (Estimated Cancer Deaths).	318960	174,520	930	0.53	86,330	49.47	25,950	14.87	61,310	35.13
Lungs & Bronchus (Men) 2014 (Estimated Cancer Deaths).	318520	173,860	930	0.53	86,000	49.47	25,860	14.87	61,070	35.13
Colon & Rectum (Men) 2010	102,740	53,160	860	1.62	25,720	48.38	8,290	15.59	18,290	34.41
Colon & Rectum (Men) 2011	101,000	50,500	820	1.62	24,430	48.38	7,870	15.58	17,380	34.42
Colon & Rectum (Men) 2012	225940	175,500	1,020	0.58	86,730	49.42	26,170	14.91	61,580	35.09
Colon & Rectum (Men) 2013	101,660	52,600	890	1.69	25,410	48.31	8,680	16.50	17,620	33.50
Colon & Rectum (Men) 2014	100620	52,540	890	1.69	25,380	48.31	8,620	16.41	17,650	33.59
Prostrate (Men) 2010	171,340	171,340	*		32,010	18.68	2,980	1.74	29,070	16.97
Prostrate (Men) 2011	170,100	170,100	*		33,680	19.80	3,130	1.84	30,590	17.98
Prostrate (Men) 2012	52,060	52,060	*		28,140	54.05	2,730	5.24	25,440	48.87
Prostrate (Men) 2013	173,590	173,590	*		29,700	17.11	2,930	1.69	26,790	15.43
Prostrate (Men) 2014	172,930	172,930	*		29,450	17.03	2,940	1.70	26,540	15.35

Graphical representation of men affected by Lung and Bronchus cancer in the US :

This graph portrays that men under 45 and above 45 have been most and least susceptible to lung and bronchus cancer , while other age groups lie in between these high and low numbers.

Figure 6.11: Death cases in Men in the US caused by Lung & Bronchus Cancer



For females:

Table 6.16: Gender and age based analysis of Death cases in Women in US

Estimated cancer Deaths Year	Cancer deaths All Sites	All sites women	Cancer deaths % female	> 45 WOMEN	% AGE >45 WOMEN	< 45 WOMEN	%AGE <45 WOMEN	>65 WOMEN	% AGE>65 WOMEN	<65 WOMEN	%AGE<65 WOMEN
2010	1,138,980	540,580	47.46	11,060	2.05	259,230	47.95	82,150	15.20	188,140	34.80
2011	1,143,900	543,040	47.47	11,110	2.05	260,410	47.95	82,530	15.20	188,990	34.80
2012	1,154,380	550,740	47.71	10,880	1.98	264,490	48.02	83,530	15.17	191,840	34.83
2013	1,151,700	546,860	47.48	10,500	1.92	262930	48.08	83180	15.21	190250	34.79
2014	1,171,440	551,420	47.07	10570	1.92	265140	48.08	83950	15.22	191760	34.78

Age distribution of Women affected by Lung and Bronchus cancer in the US is shown below:

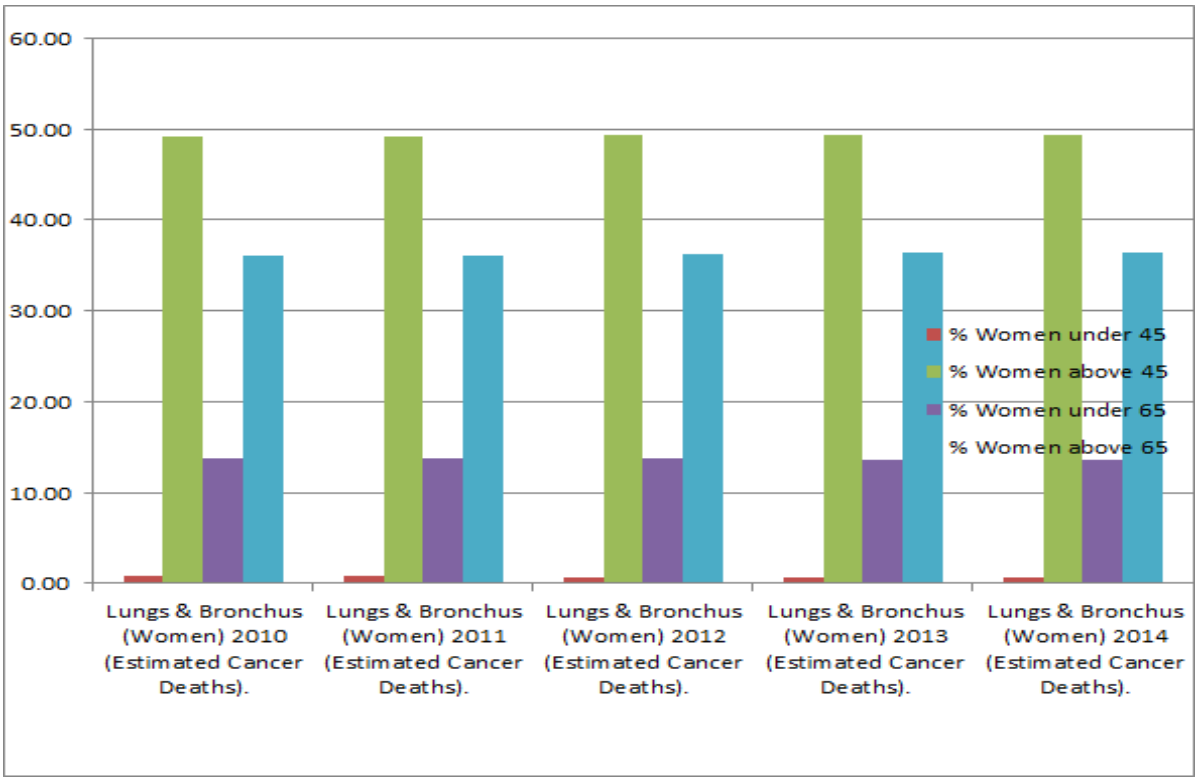
Table 6.17: Age distribution of Women affected by Lung & Bronchus Cancer in US

Prevalent Death cancer in a given year	Total Death Cancer cases in US	Women of all ages	Women under 45	% Women under 45	Women above 45	% Women above 45	Women under 65	% Women under 65	Women above 65	% Women above 65
Lungs & Bronchus (Women) 2010 (Estimated Cancer Deaths).	314,600	142,160	1,110	0.78	69,970	49.22	19,710	13.86	51,370	36.14
Lungs & Bronchus (Women) 2011 (Estimated Cancer Deaths).	313,880	142,680	1,120	0.78	70,220	49.22	19,780	13.86	51,560	36.14
Lungs & Bronchus (Women) 2012 (Estimated Cancer Deaths).	198,120	145,180	1,010	0.70	71,580	49.30	19,910	13.71	52,680	36.29
Lungs & Bronchus (Women) 2013 (Estimated Cancer Deaths).	318,960	144,440	910	0.63	71,310	49.37	19,610	13.58	52,610	36.42
Lungs & Bronchus (Women) 2014 (Estimated Cancer Deaths).	318,520	144,660	930	0.64	71,400	49.36	19,680	13.60	52,650	36.40
Colon & Rectum (Women) 2010	102,740	49,580	710	1.43	24,080	48.57	5,860	11.82	18,930	38.18
Colon & Rectum (Women) 2011	101,000	50,500	820	1.62	24,430	48.38	7,870	15.58	17,380	34.42
Colon & Rectum (Women) 2012	225,940	50,440	740	1.47	24,480	48.53	6,160	12.21	19,060	37.79
Colon & Rectum (Women) 2013	101,660	49,060	720	1.47	23,810	48.53	6,170	12.58	18,360	37.42
Colon & Rectum (Women) 2014	100,620	48,080	700	1.46	23,340	48.54	6,040	12.56	18,000	37.44
Breast (WoMen) 2010	79,751	79,680	2,640	3.31	37,200	46.69	17,030	21.37	22,810	28.63
Breast (WoMen) 2011	79,111	79,040	2,620	3.31	36,900	46.69	16,890	21.37	22,630	28.63
Breast (WoMen) 2012	79,091	79,020	2,510	3.18	37,000	46.82	16,790	21.25	22,720	28.75
Breast (WoMen) 2013	79,311	79,240	2,460	3.10	37,160	46.90	16,770	21.16	22,850	28.84
Breast (WoMen) 2014	69,542	69,480	2,480	3.57	37,520	54.00	2,940	4.23	26,540	38.20

Graphical representation of women affected by Lung and Bronchus cancer in the US :

This graph portrays that women under 45 and above 45 have been most and least susceptible to lung and bronchus cancer , while other age groups lie in between these high and low numbers.

Figure 6.12: Death cases in Women in the US caused by Lung & Bronchus Cancer



7. Conclusions

This project has been about taking data from an open source database , which is cancer.org and analyse this data to make reliable conclusions . The number of men and women affected by different types of cancer in years 2010,2011,2012,2013 and the prediction for 2014 is analysed in the process. The population for the different states was calculated using a mathematical formulae . Then the predictions and the death cases were compared against this data. The conclusion we could arrive at were that california was not the most effect but Maine. Also the most fatal cancer amongst men is Prostate cancer and the most fatal for women the last years the comings years is Breast cancer. The types of cancers to be in a lookout for is Breast cancer, Prostate Cancer, Colon Cancer & Rectum Cancer. More resources have to available to treat the above mentioned types of cancer. As for people who would like to further understand the implications and read into more detail about the same, you could refer the Bibliography and there is lot more data to work with and also understand in detail about these types of cancers and their implications.

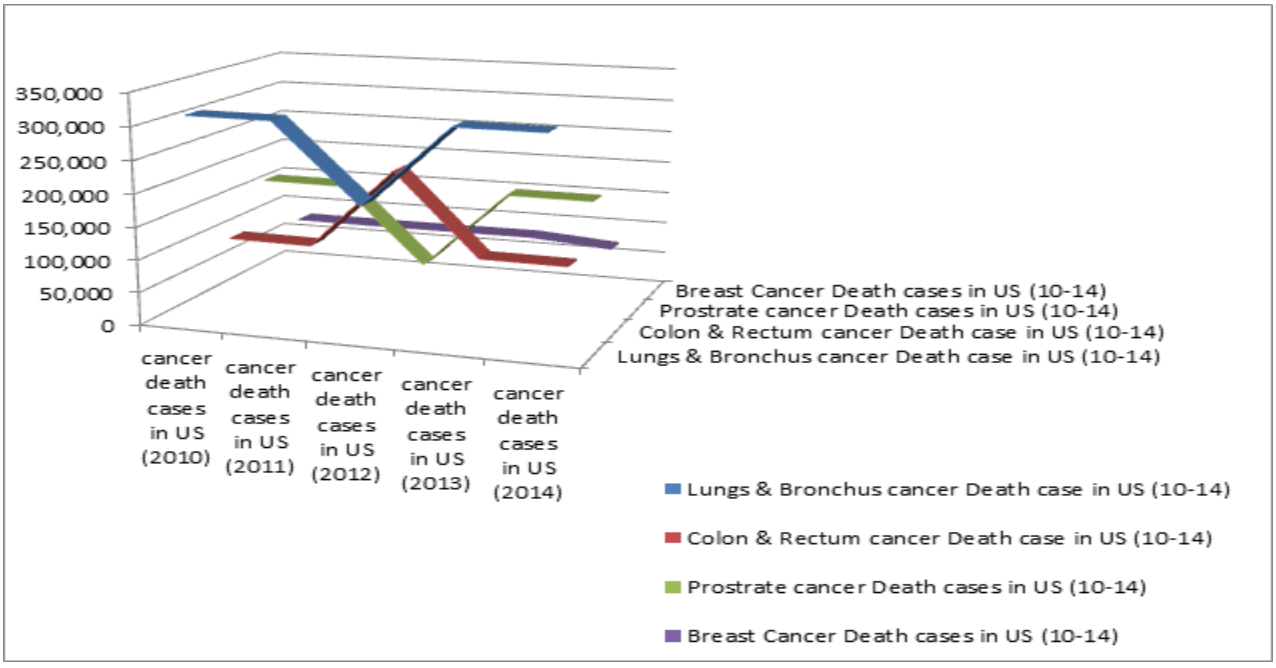


Figure 7.1: Conclusion - Total Death cases caused by Lung & Bronchus,Colon , Prostate and Breast Cancer across US

8. Summary

One of the challenges that we faced on this project is we couldn't find a state by state population for the current year (2014). According to census.gov, the data will be scheduled for release on December 2014. Since we need that data to derive ratio, we came up with our own estimate, using 2013 state population and add the current growth rate.
(http://en.wikipedia.org/wiki/List_of_U.S._states_by_population_growth_rate)

Another challenge is that we could not come up with the data set we want to study because most of the data or the case number are just estimate. But professor assured as that it is alright to use the data set in cancer.org to make complete this study , and indeed, we were able to derive the solution that we wanted.

Lesson we learned in doing our project is time management. No matter how simultaneous our project with our other courses, we find time to collaborate and finish our project in time.

The next step for this project is to use sophisticated tools to further enhance the analysis we had on this project, because we only used MS Excel for most of our data. Another thing is if the whole group will be enrolled on the Advance Data Science class on the fall, maybe Professor can lead us to better data science analysis using the same project.

9. Recognition/Comments

Since we are beginners to Data Science project, it is very hard to come up with a project without a based project since we are doing a pilot project school wise. Though we have step by step guidance with our Professor, we still lack this aspect. The advantage of having this though is we are free to do what we think is the best way for our analysis to solve our problem.

An improvement that we can think of for this project in the future is the use of database, even the easiest database like MS Access. As mentioned earlier in Methodology (part 4), the group took time settling for a set of data to use that ate all our time and resources that could have place our data in a decent database at least.

10. Project Schedule

Schedule/Timetable:

Following are the basic weekly milestones for the progress as discussed during project start-up:

Week 1 (January 15): Class start, group assignment.

Week 2 (January 22): Each member coming up with 3 topics each.

Week 3 (January 29): Brainstorming which topic to do for the project.

Week 4 (February 5): Finalize project topic which is Cancer Statistics in the United States.

Week 5 (February 12): Started gathering data from Cancer.org and Created our project template.

Week 6 (February 19): Worked on project documents.

Week 7 (February 26): Worked on project documents.

Week 8 (March 5): Midterm. Project Milestone Check.

Week 9 (March 12): Discussion with Data Analysis and dissemination of tasks.

Week 10 (March 19): Discussion on Data Analysis and graphs.

Week 11 (March 26): Discussion on final project documentation.

Week 12 (April 2): Project Finalization.

Week 13 (April 9): Project Presentation

Week 14 (April 16): Project Presentation for our group.

Week 15 (April 23): Final

11. Bibliography

"Big Data Giant Joins InfoChimps to Save the World's Structured Information." *ReadWrite*. N.p., n.d. Web. 06 Mar. 2014. <http://readwrite.com/2011/01/03/data_giant_climbs_aboard_at_infochimps#awesm=~oAI0wRqHOiji0n>.

"Cancer Facts & Figures 2010." *American Cancer Society*. N.p., n.d. Web. 10 Feb. 2014. <http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2010/index>.

"Cancer Facts & Figures 2011." *Cancer Fact and Statistics 2011*. N.p., n.d. Web. 11 Feb. 2014. <http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2011/index>.

"Cancer Facts & Figures 2012." *American Cancer Society*. N.p., n.d. Web. 14 Feb. 2014. <http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2012/index>.

"Cancer Facts & Figures 2013." *American Cancer Society*. N.p., n.d. Web. 30 Jan. 2014. <http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2013/index>.

"Cancer Facts & Figures 2014." *Cancer Facts & Figures 2014*. N.p., n.d. Web. 16 Feb. 2014. <http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index>.

"Census Bureau Homepage." *Census Bureau Homepage*. N.p., n.d. Web. 02 Feb. 2014. <https://www.census.gov/>.

Green, John. *The Fault in Our Stars*. New York: Dutton, 2012. Print.

"List of U.S. States by Population Growth Rate." *Wikipedia*. Wikimedia Foundation, 04 May 2014. Web. 08 Mar. 2014. <http://en.wikipedia.org/wiki/List_of_U.S._states_by_population_growth_rate>.

12. Appendix I

13. Glossary :

American cancer society: Established since 1931, is a health organization dedicated to eliminate cancer throughout the United States.

Cancer: The disease caused by malignant and invasive growth of tumor in a part of body.

Cancer registry: Systematic collection of data about cancer and tumor diseases.

Data: The facts and statistics collected together for reference or analysis.

Database: The organised collection of data.

Data mining: The computational process of discovering pattern in large data sets involving methods at the intersection of machine learning, statistics and database.

Data Science: The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it.

Data warehouse: Used for reporting and data analysis. It involves integrating data from more than one disparate sources creating data warehouse.

Death Cases: The collection of reported cancer death cases from different source (treatment facilities, clinics, pathologists, death certificates) registered every year in well defined populations.

Deductive argument: The process of reasoning from one or more general statements to reach a logically certain conclusion.

Hypothesis: A proposed explanation for a phenomenon.

Leukemia: Type of blood or bone marrow cancer. It is invasive growth of immature white blood cells.

Lungs and Bronchus cancer: It is a lung tumor characterised by uncontrolled cell growth in tissues of lungs.

New cancer cases: The collection of new reported cancer cases from different sources (treatment facilities, clinics, pathologists, death certificates) registered every year in well defined populations.

Prostate cancer: Cancer developed in the prostate, a gland in the male reproductive system

Relational database: It is a collection of tables of data items, formally described and organised according to the relational model.