Introduction and Motivation
Cancer is the name of a group of diseases that affects a large percentage of the world population. Only in the United States in 2012, it is estimated as 1,638,910 new cases and 577,190 deaths [1]. The goal of this study is to analyze cancer occurrences and their relation with environmental condition. Along with the cancer data, we used the air quality data (air quality of each state) and the census data (business patterns in each state).

Dataset
- Cancer: United States Cancer Statistics from 1999 and 2008. Contains the rate (per 100,000 population) of incidence and mortality in each state. It is grouped by the type of cancer, sex, age, and race. [2]
- Area of all the states in US. [5]

Our dataset consists of 465,300 rows of items each with 127 attributes. We assume that the business patterns and air quality levels in each state are similar over the years.
Findings

1. Men are more likely to have cancer than women and the gap is bigger in African American/Black because they have the highest incidence rate of prostate cancer.

An age-adjusted ratio is used in order to compare different groups with a priori different age distributions. The cancer disease is highly correlated with the age so comparing a younger population with an older population would not be fair using a crude ratio. See [6] for a precise definition of the age adjustment.

Figure 1a: Rates (per 100,000 persons) of incidence and mortality by sex and race along the period of time 1999-2008.
Figure 1b: Differences of incidence rates between sex.

The difference between sex is explained mainly by the difference between rates of prostate cancer and breast cancer. In the line chart, cancers are colored (i.e., red or blue) if they only appear in a single type of gender (i.e., female or male) and not both. The sex dependence is shown in the pie chart.
2. Female and Male have different distributions of incidence count over age groups. Female curve rises earlier but male curve goes higher.

Figure 2a: Average incidence for all age groups, one line to each sex.

This figure shows two line graphs each for male and female. We can notice several interesting findings here. First, we see that the female curve rises in an earlier state than the male curve. Second, we can see that once the male curve starts to rise, it grows higher than the female curve. We got curious of what kind of cancer creates this distribution and plotted another visualization to easily capture this information.
Figure 2b: Average incidence for all types of cancer binned by age group.

This figure is similar to Figure 2a but divided each line by “types of cancer”. We can notice that different cancer types have different distribution over age and the ones that provoke the curves are “Female Breast” and “Prostate”. It was interesting to see that females have a earlier burst than male because of breast cancer. Therefore, females should start to be careful in their early age to prevent breast cancer.
3. Incidence and mortality rates are decreasing over the years. Only “Liver and Intrahepatic Bile Duct” cancer mortality is increasing.

Figure 3a: The linear regression of the mean rates shows that the number of cancers per 100,000 is decreasing.
Figure 3b: This bar chart represents the slopes of the regression lines that model the mean rates variation over time by type of cancer. Only “Liver and Intrahepatic Bile Duct” cancer mortality is increasing.
4. Not all the states are evolving in the same way.

Figure 4a: The slope of the regression line reveals that there are different evolutions in the cancer rates along the period 1999-2008. The Y-axis is the slope of the regression line.

The incidence and mortality rates are decreasing in almost all the states, but there are some exceptions. Moreover, we found that the speed of change is not homogeneous between types of cancer within a state, so there are particular types of cancer, like the "Prostate" cancer in Alaska, that change 3 or 4 times faster than the mean. The comparison between Hawaii, which is the first in increasing rates, and Alaska, which is the first in decreasing rates, shows this reasoning. (See Figure 4b)
Figure 4b: This chart shows the evolution of all types of cancer in Alaska and Hawaii. The colored lines are the cancers that makes the difference.
5. Different age groups are sensitive to different types of cancer.

In this figure, we see that more likely cancer differs among age groups. For example, really young people (i.e., under age 14) have to be careful of "Leukemias," "Brain and Other Nerve System." For people in the 15-29 age frame, "Lymphomas" becomes critical. In the thirties and forties, "Female Breast Cancer" is really critical and "Digestive System" arises. In the 55-75 age frame, "Prostate" and "Lung and Bronchus" appear in the top three cancers. After age 75, "Digestive System" becomes dominant.

Figure 5: Incidence sum of each cancer type for all age groups.
6. Cancer rates are more correlated with business patterns than with air pollution.

Figure 6: Correlation between “age adjusted rate” and census and air pollution. The highlighted elements are the ones that have a significant correlation.

This figure shows four significant relationships. By using Spotfire’s “Data Relationships” tool, we spotted there are several interesting correlations between “age adjusted rate” and the census data. We used the linear regression model to analyze the data and found several negative correlations ("arts, entertainment, and recreation," "information," and "real estate and rental and leasing") and positive correlations ("manufacturing", "healthcare and social assistance"). This results were surprising because we expected both groups (air pollutants and business activities) to have some correlation with the age adjusted rates, but we only found a significant linear correlation (p-value < 0.01) with business patterns.
Tool Critiques

By exploring several tasks with a massive dataset, we found that both systems include different strength and weakness, but generally they enable users to explore the data much easier than programmatically or with spreadsheets tools. Also by interactively re-visualizing the data, users can get more insight out of the data. Here, we’ll criticize both systems based on our observations. Notice that we have used SpotFire more than Tableau because we found it easier and more effective.

SpotFire

Positive aspects:

- The paradigm based on standard charts is easier to understand and faster for creating linked views.
- Bookmarks are a good solution for saving visualization states so they can be loaded easily in the future. Furthermore, they support a basic way of collaboration between analyzers.
- Easy to join multiple tables
- Consistent color scheme is used for same categorical attribute
- Easy to add an additional computed column using built-in functions
- Handling missing values are well done
- Trellis support is very effective for comparisons

Limitations and suggestions:

- Adjusting font size and re-positioning labels is time consuming. It would have been nice if this could be done automatically so that all the fonts are visible with no overlaps.
- No fast way of visualize the difference of different columns. Now it is done by adding a new column but a direct access would be better.
- Hard to reorder x-axis or y-axis labels the way users need it

Tableau

Positive aspects:

- The paradigm of visual grammar that tableau uses allows more flexibility while designing a particular visualization
- Incorporates maps to create fast cartographic charts
- Easy to sort and rearrange marks and axes
- Very good support of labeling with easy re-positioning
- Selections do not change the color like in Spotfire but desaturate the non selected elements.

Limitations and suggestions:

- At first it was not easy to understand how they automatically separated “dimensions” and “measures”.
- Not easy to create linked views

Conclusion

We presented six findings based on visual evidences. All the work was done with SpotFire and Tableau. We tried to use Hierarchical Clustering Explorer (HCE), but it did not work for our dataset since it was too large.
Currently, our study is limited to few environmental factors (i.e., air pollution and business patterns in each state), but more interesting attributes can be easily added. In addition, currently our cancer data is grouped by state-level. If all the attributes can be grouped by county-level, we assume it will provide more accurate and meaningful results.

Reference