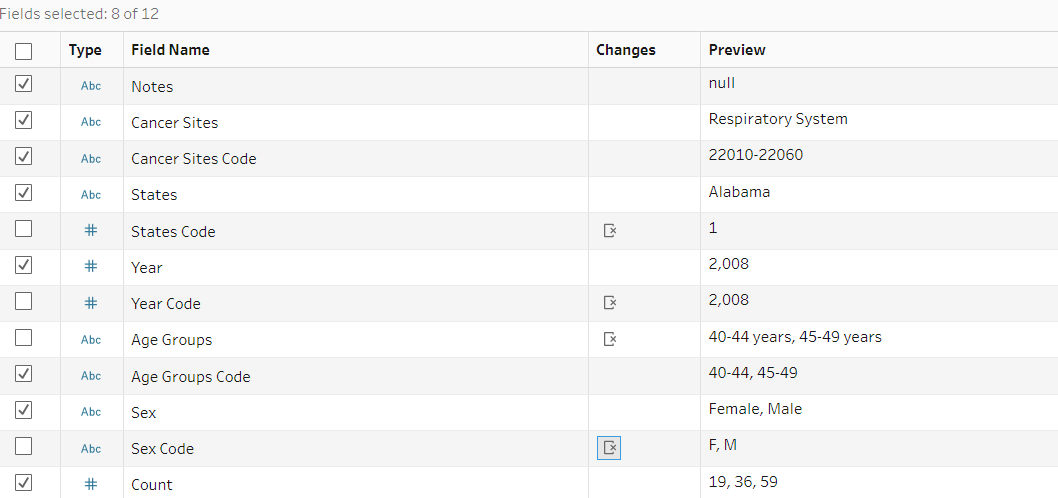
**Visual Analytics Final Project**

Janna Schlageter, Ahmad Alfahad, Diaa Juneidi, Ali Al Rawendoozi, and Sohil Bhavsar

**Chapter 2: Filtering**

During close examination of data, we found there are some data column fields which are repeating. We use filtering techniques at the data input step to remove unnecessary or repeating fields as below. We filter State Code as for our visualization we need state name only. State code is not required. Then we filter year code as year and year code are similar columns. Then we filter Age groups as Age groups and Age Groups code has only difference of string “years”. Then we filter sex code to keep only one field which states female & male patients.



**Chapter 2 - Filter by Columns**

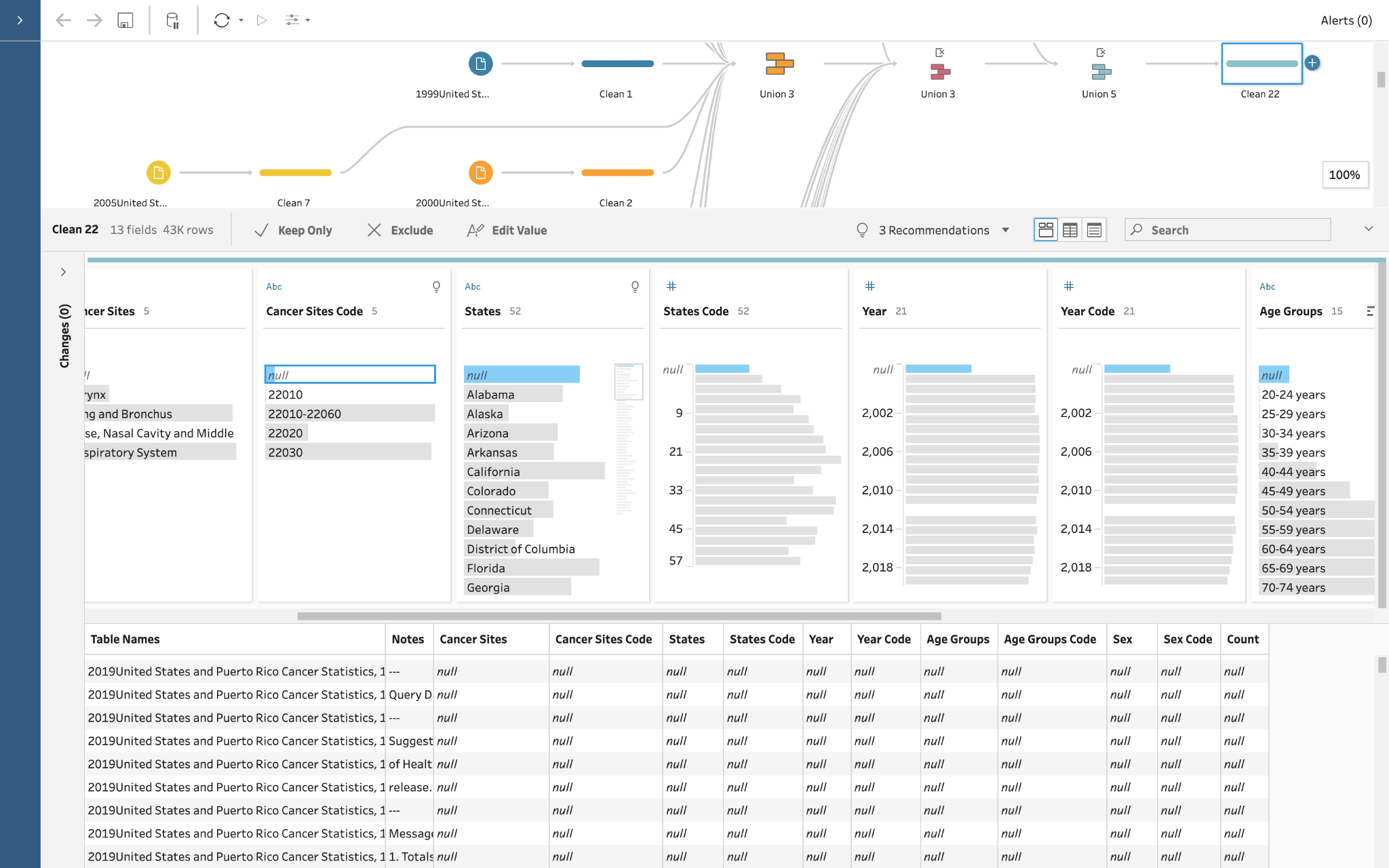
Filtering by columns in Tableau is a powerful tool that allows you to focus on specific subsets of data within a larger dataset. To filter by columns in Tableau, you first need to select the column that you want to filter. You can do this by clicking on the column header in the view, or by dragging the column to the Filters shelf.

Once you have selected the column you want to filter, you can apply a filter by clicking on the drop-down menu in the Filters shelf and choosing the type of filter you want to apply. Tableau offers a variety of filtering options, including basic filters, conditional filters, and top and bottom filters.

Basic filters allow you to filter data based on a single criterion, such as a specific value or range of values. Conditional filters allow you to create more complex filters based on multiple criteria, such as filtering for data that falls within a certain date range or meets specific conditions.

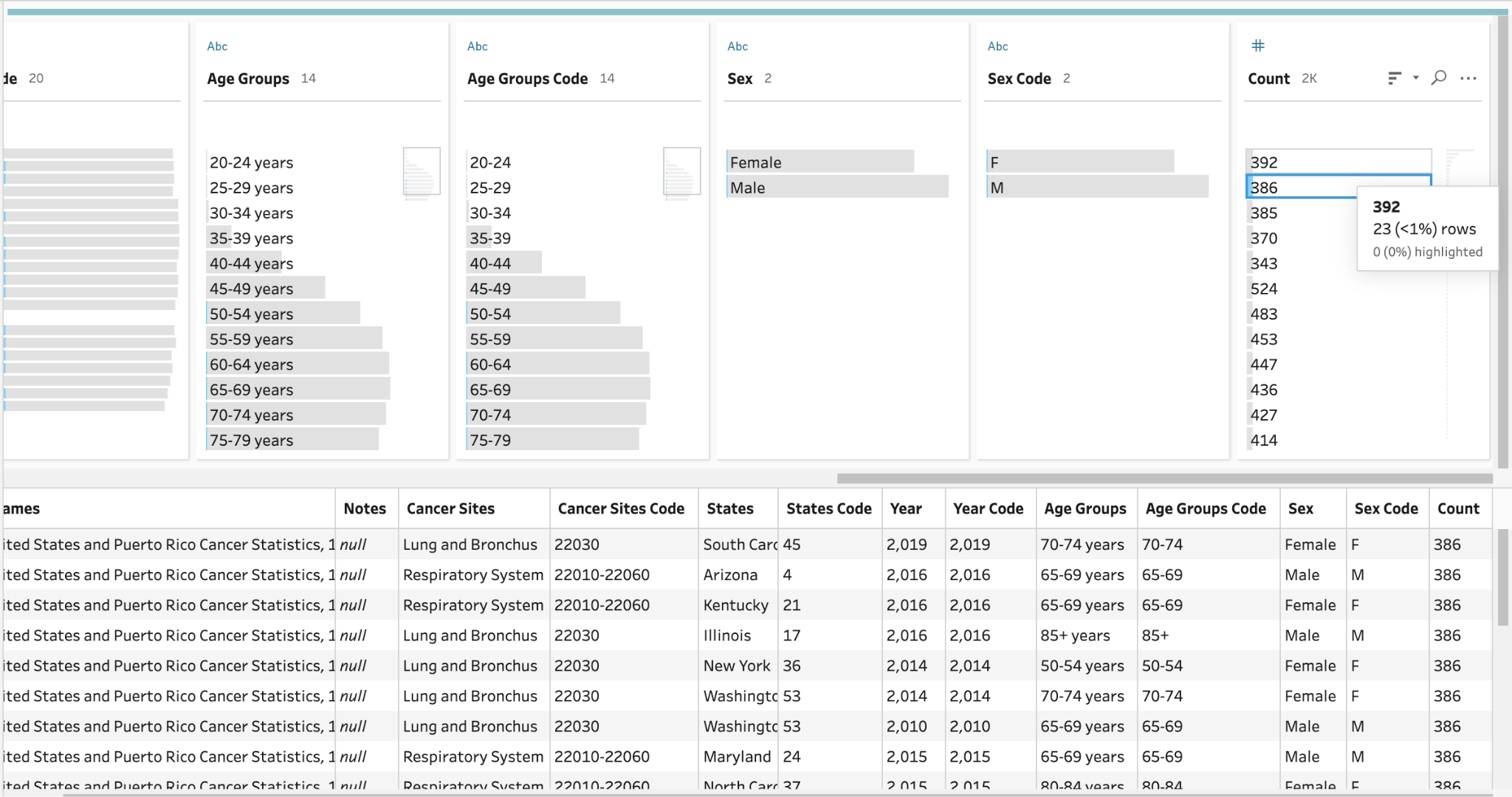
Top and bottom filters allow you to filter for the highest or lowest values in a dataset, based on a specific measure. For example, you could use a top filter to show the top 10 products by sales, or a bottom filter to show the 5 least profitable regions.

Overall, filtering by columns in Tableau is a powerful way to focus on specific subsets of data within a larger dataset, and can help you gain insights and make more informed decisions based on your data.



**Chapter 2 : Explore the Data to Identify Patterns**

To further understand and interpret the data, we explored the data. In the cleaning step we were able to sort and filter the columns to view different instances. We were able to sort the ages column to view the instances according to the age groups, this gave quick insight to which ages have higher rates of cancer. We were also able to sort the count of instances descending in order to view the states, years, and ages with the highest counts of cancer. This will help us understand the data and create visualizations to draw conclusions.



**Chapter 3 : Sort**

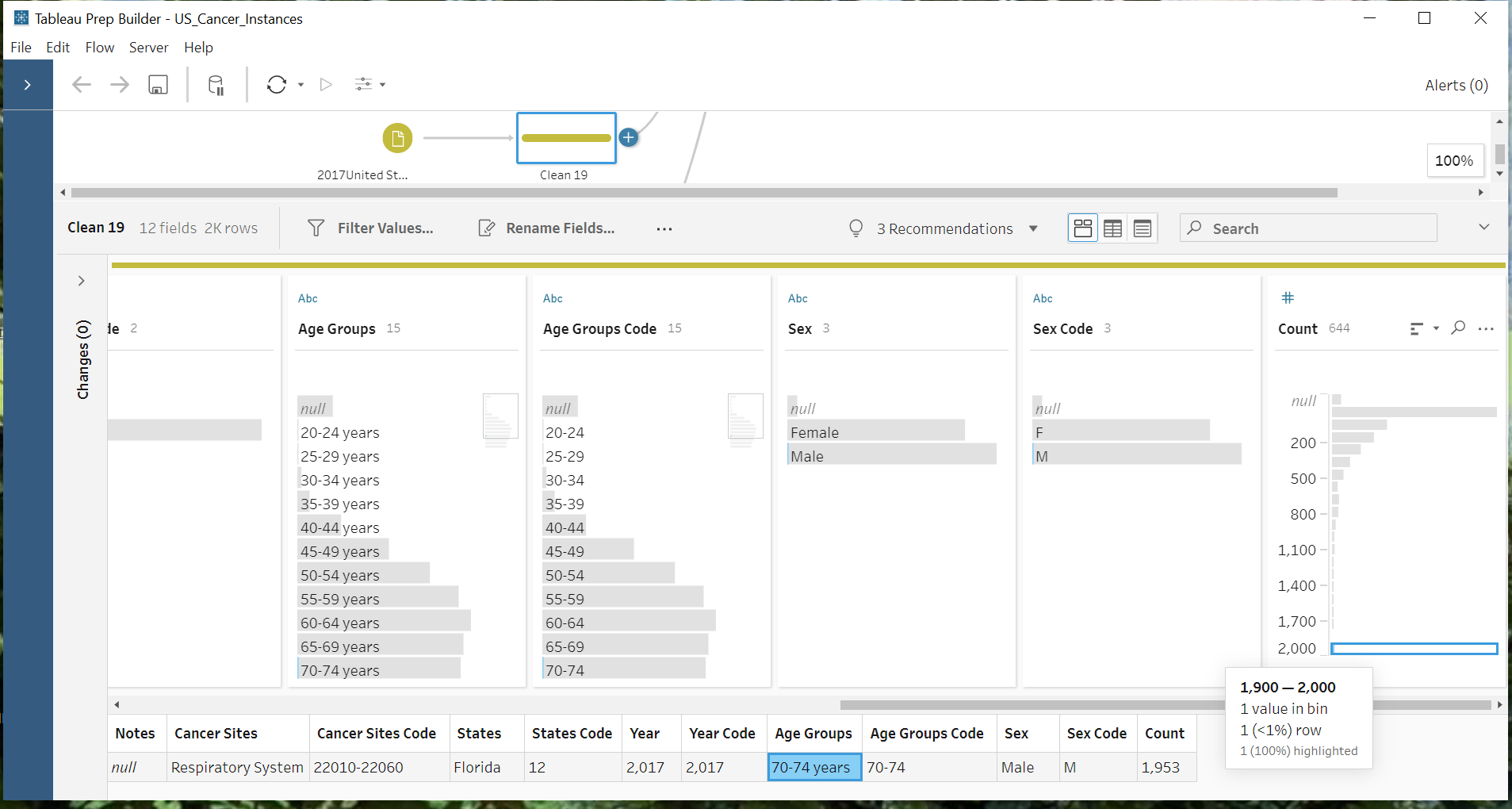
After filtering we add clean steps to each dataset. At the clean step first we use sort technique for count field to visualize data of approx. number of patient count which is common for the patients related to Lung cancer with respect to other data fields. For example, per below fig. We can easily visualize here that for 1999 the most number of patient counts were 16 patients identified related to Lung cancer with respect to different regions and age bands. We use sort technique in descending (Highest to Lowest count).





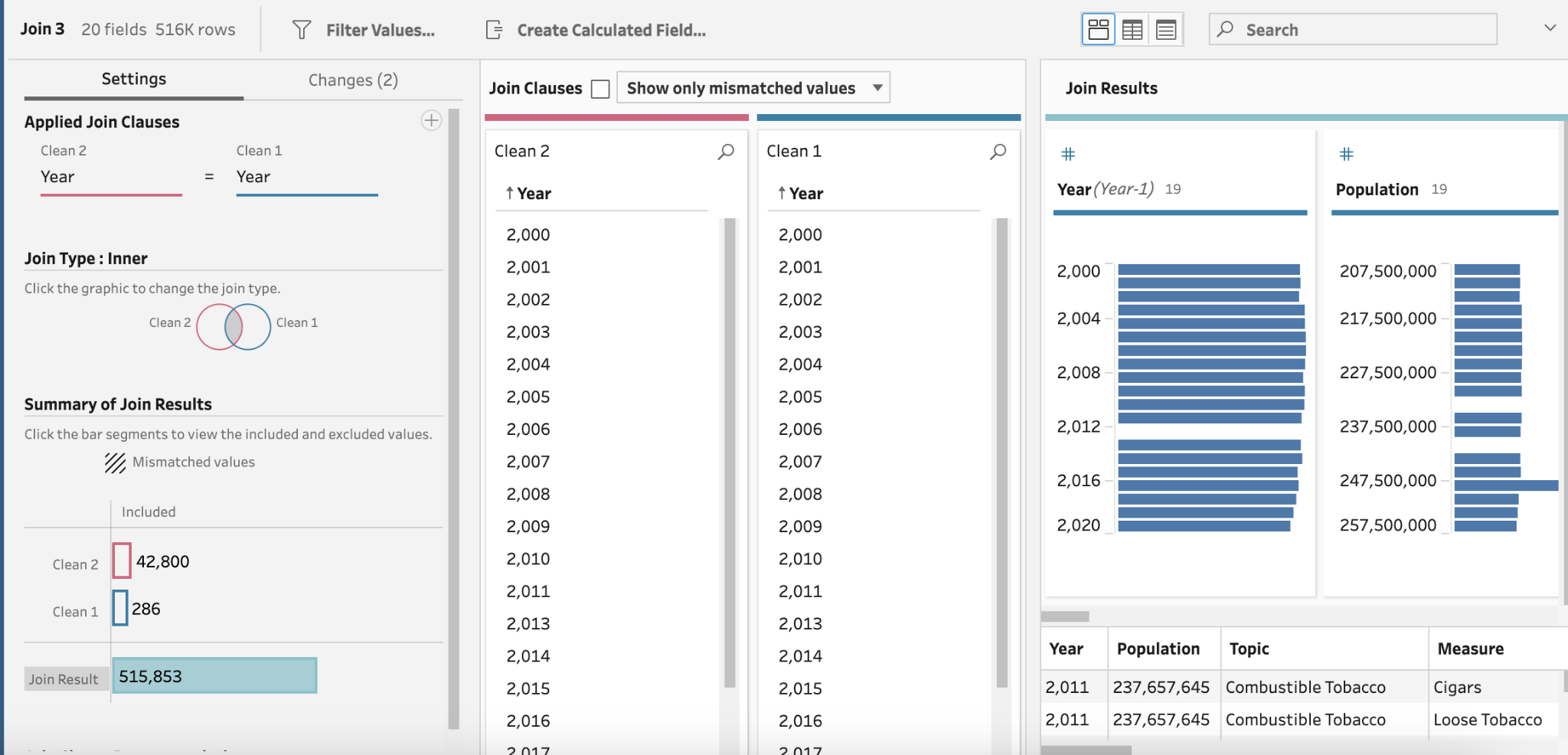
**Chapter 3: Outliers**

One of the most important steps in the cleaning process is to use sorting techniques to identify any outliers in our dataset. During this process, we noticed an outlier in the count of cases for adult men his age between (70-74) in Florida. We decided to investigate this outlier further to determine whether we should remove it from the dataset. After conducting some research, we discovered that Florida is one of the states with the highest rates of tobacco use among adults, with 40% of adults using cigarettes this year . Based on this information,we concluded that the outlier was a genuine case and should not be removed.



**Chapter 3: Join Data**

To create visualization involving both tobacco rates and cancer rates, we used the join feature to combine the two datasets. We used a center join which only used data that was contained in both of the datasets. We set the two year columns equal to each other in the join. Since the tobacco rates did not include data from the year 1999, the join only combined data between 2000 and 2019.



**Chapter 4: Missing Data**

Handling missing data in Tableau is an important part of the data analysis process. There are several ways to handle missing data in Tableau, depending on the nature of the data and the analysis you are performing.

One approach to handling missing data is to simply exclude it from your analysis. This can be done by filtering out any rows or columns that contain missing data, or by using the "Exclude" option in the data source tab.

Another approach is to replace missing data with a default value, such as zero or the average value of the column. This can be done using calculated fields or data blending techniques.

A more advanced approach to handling missing data is to use imputation techniques to estimate missing values based on the values of other variables in the dataset. This can be done using statistical models or machine learning algorithms, and can be a powerful way to fill in missing data and improve the accuracy of your analysis.

Regardless of the approach you choose, it's important to carefully consider the impact of missing data on your analysis and to choose the best approach for your specific situation. By handling missing data effectively, you can ensure that your analysis is accurate and reliable, and that you are making informed decisions based on the best available data.

**Chapter 4 - processing data**

Processing data in Tableau involves several steps, including connecting to data sources, cleaning and preparing the data, and creating visualizations and dashboards.

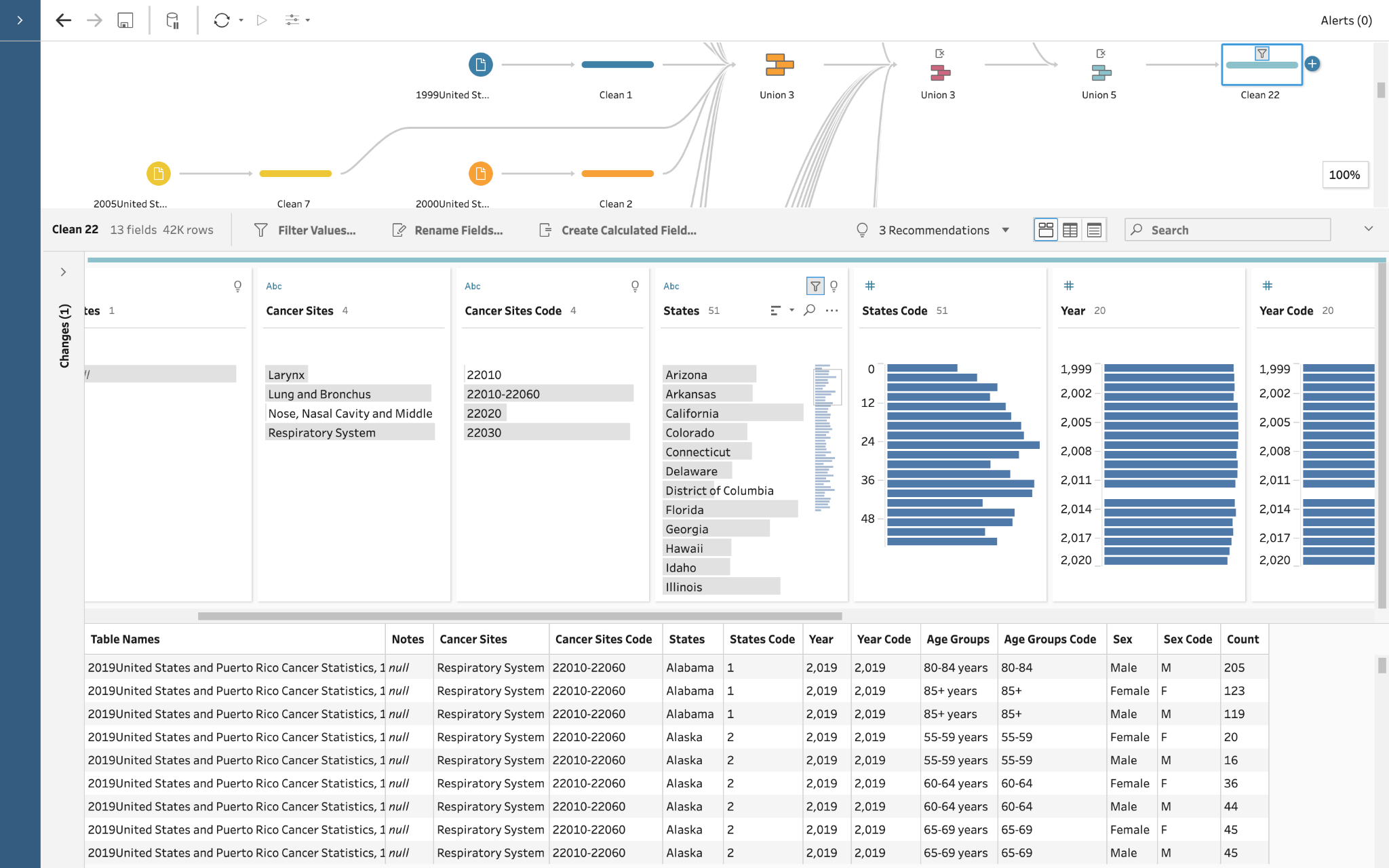
The first step in processing data is to connect to your data source. Tableau supports a wide variety of data sources, including spreadsheets, databases, and cloud-based services. Once you have connected to your data source, you can begin cleaning and preparing the data for analysis.

Cleaning and preparing data involves several tasks, such as removing duplicates, correcting errors, and transforming data into a format that is suitable for analysis. Tableau offers several tools to help you clean and prepare your data, such as data blending, data reshaping, and data cleaning functions.

Once your data is cleaned and prepared, you can begin creating visualizations and dashboards. Tableau offers a wide variety of visualization types, such as bar charts, scatter plots, and heat maps, as well as advanced features such as calculated fields, filters, and parameters.

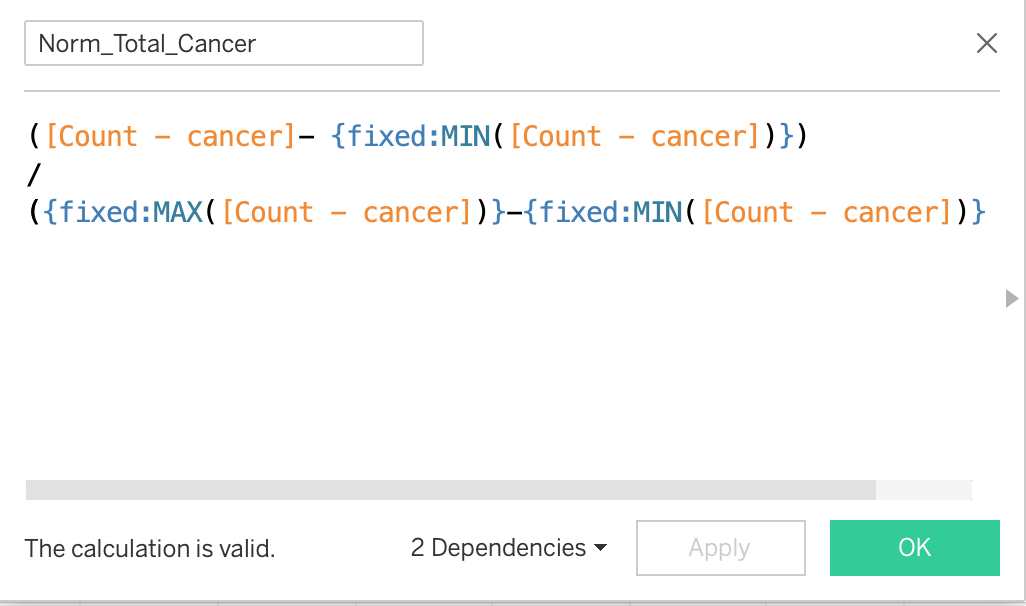
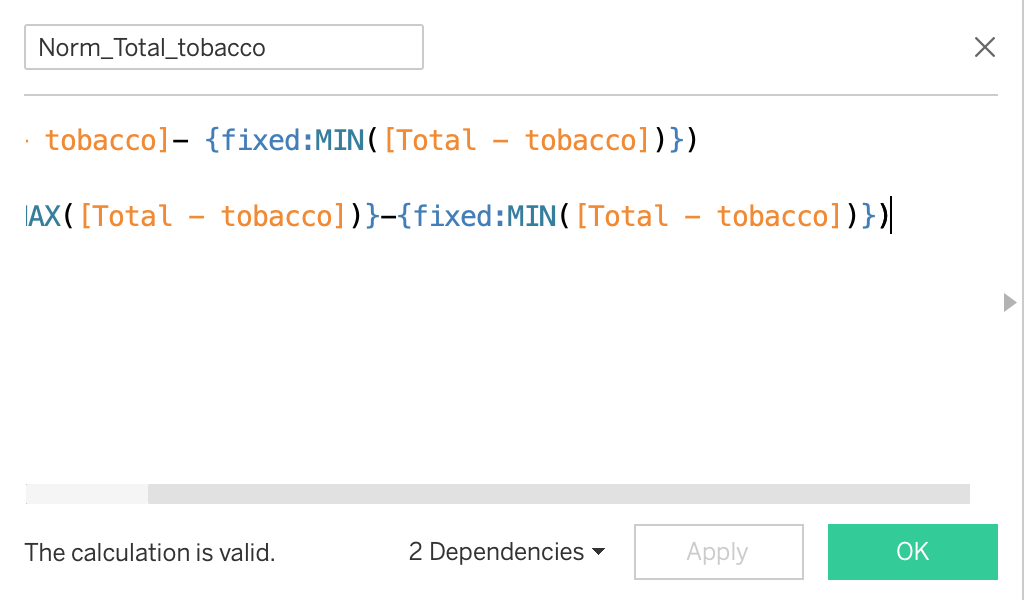
Creating effective visualizations and dashboards involves careful consideration of the data and the audience for your analysis. You should choose the most appropriate visualization types for your data, and design your dashboards to be clear, concise, and easy to understand.

Overall, processing data in Tableau involves several steps, from connecting to data sources to creating visualizations and dashboards. By following best practices for data cleaning and visualization design, you can create accurate, informative, and impactful analyses that help you make better decisions based on your data.

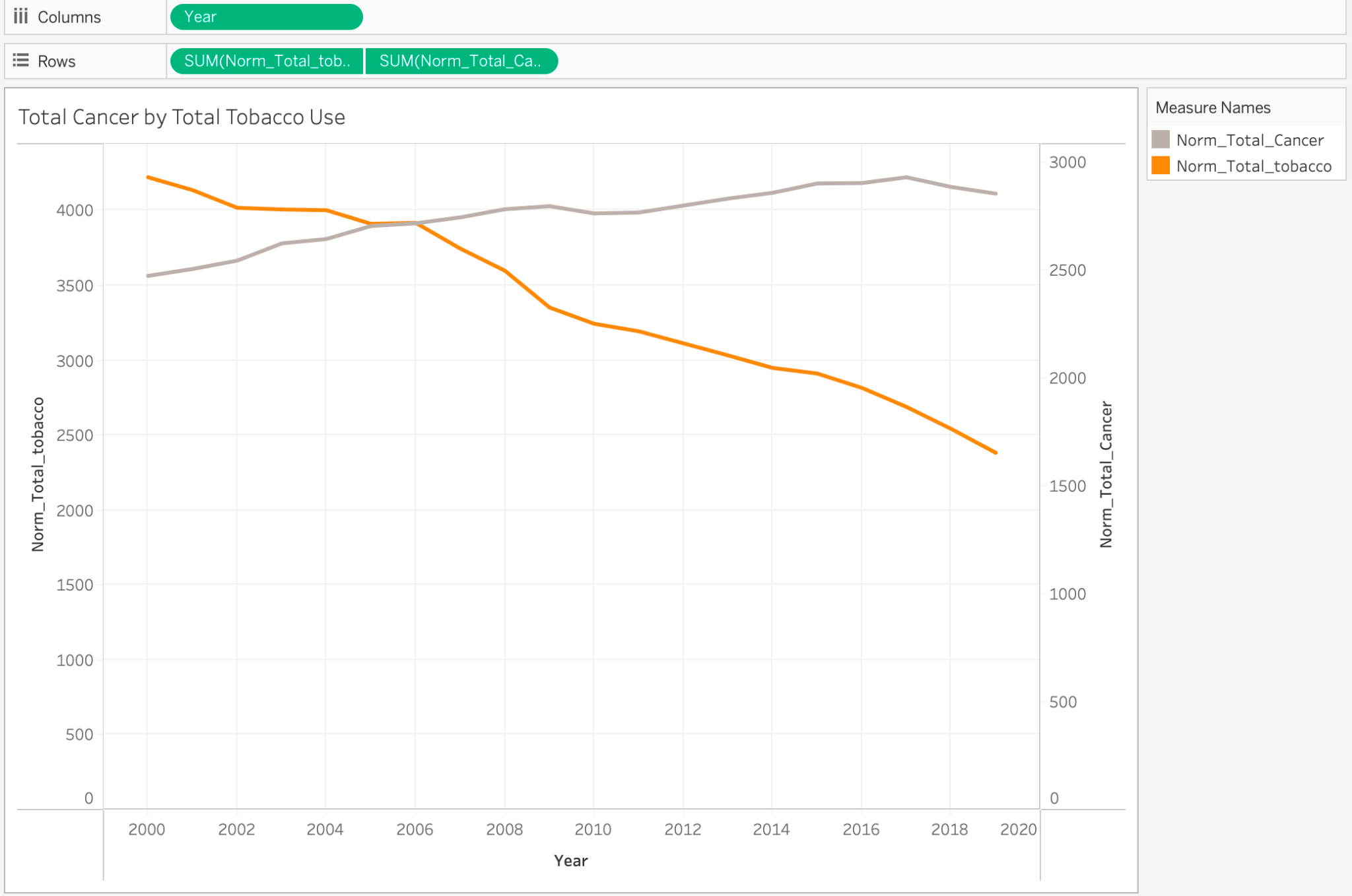


**Chapter 4: Normalization**

In order to properly compare the rates of tobacco use and cancer instances over time, we will normalize this data. Since the two rates are on very different scales, this will make it easier to interpret the changes over time in both. To do this we created a calculated field in Tableau and used the normalized equation. These can be seen in the following screenshots.



Then with these normalized fields, we were able to plot these over time and compare the two rates with similar scales.



**Chapter 5: Sequence/Graph**

Sequencing and graphing are important parts of data analysis, and Tableau provides several tools to help you sequence and graph your data effectively.

Sequencing involves organizing your data in a logical order, such as by time, category, or value. Tableau provides several ways to sequence your data, such as by using sorting, grouping, and filtering functions. For example, you can sort your data by date, group your data by region, or filter your data to show only certain categories.

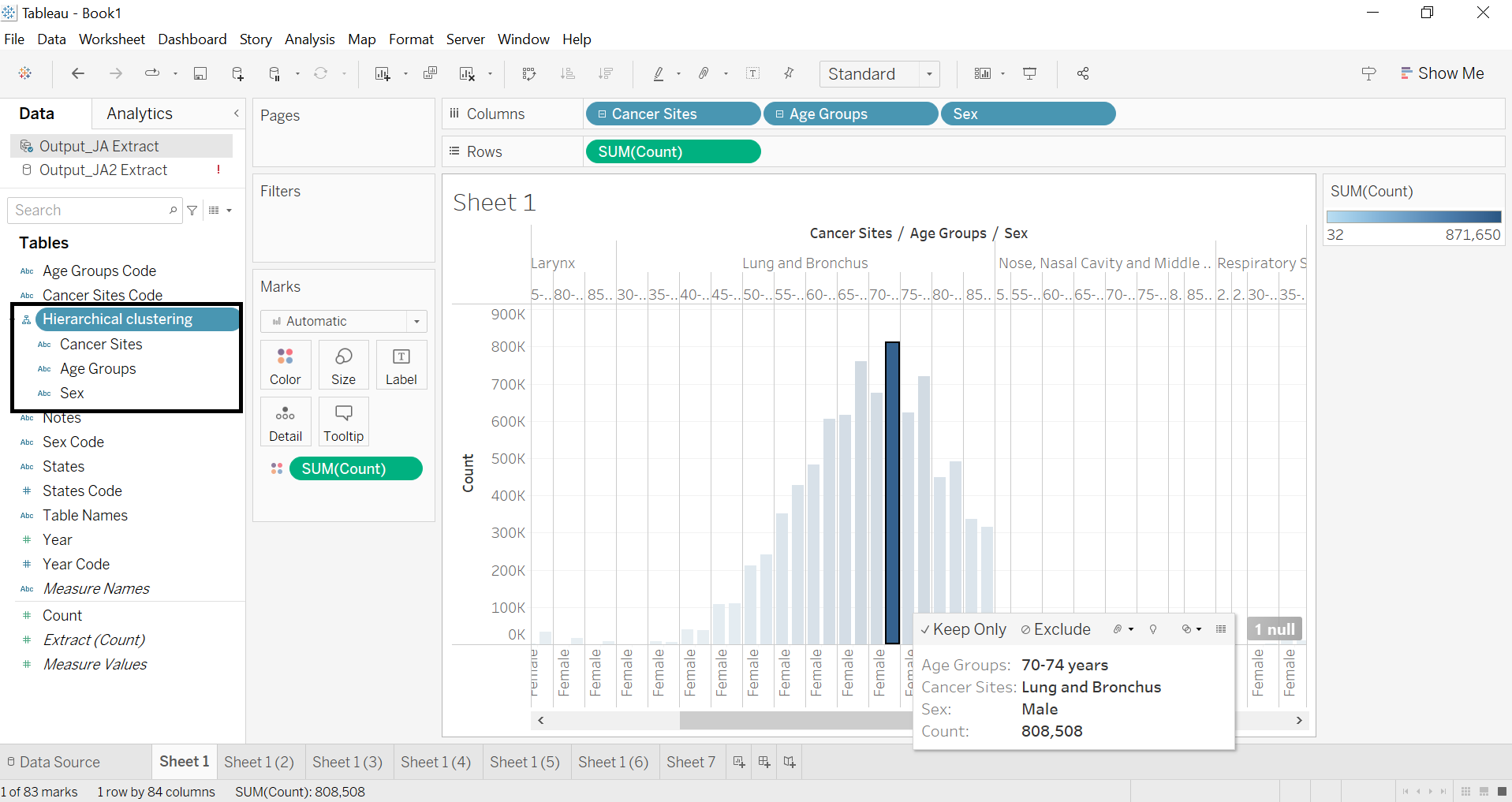
Graphing involves creating visual representations of your data, such as charts, graphs, and maps. Tableau provides a wide variety of graph types to choose from, such as bar charts, line charts, scatter plots, and heat maps. You can also customize your graphs by adding labels, colors, and other visual elements.

To create effective graphs in Tableau, it's important to choose the most appropriate graph type for your data, and to design your graphs to be clear, concise, and easy to understand. You should also consider the audience for your analysis, and design your graphs to be visually appealing and engaging.

Overall, sequencing and graphing are essential parts of data analysis, and Tableau provides several tools to help you sequence and graph your data effectively. By following best practices for sequencing and graphing, you can create informative, accurate, and impactful analyses that help you make better decisions based on your data.

**Chapter 5 –Hierarchical clustering:**

The hierarchical clustering technique helped us to explore the dataset in greater depth and discover new visualizations for the data. By combining cancer sites, age group, and sex into one group, this technique provided us with new insights from multiple perspectives. Using the drill-down hierarchical approach, we were able to delve deeper into the data and obtain more detailed information. Through this process, we found that the highest number of lung cancer cases occurred in males aged between 70 and 74, with a total of 808,508 cases.



**Chapter 5 : Density Base**

Density-based analysis is a way to analyze data that is related to a geographic location, such as population density or crime rates. To create a density-based analysis in Tableau, you need to follow a few steps.

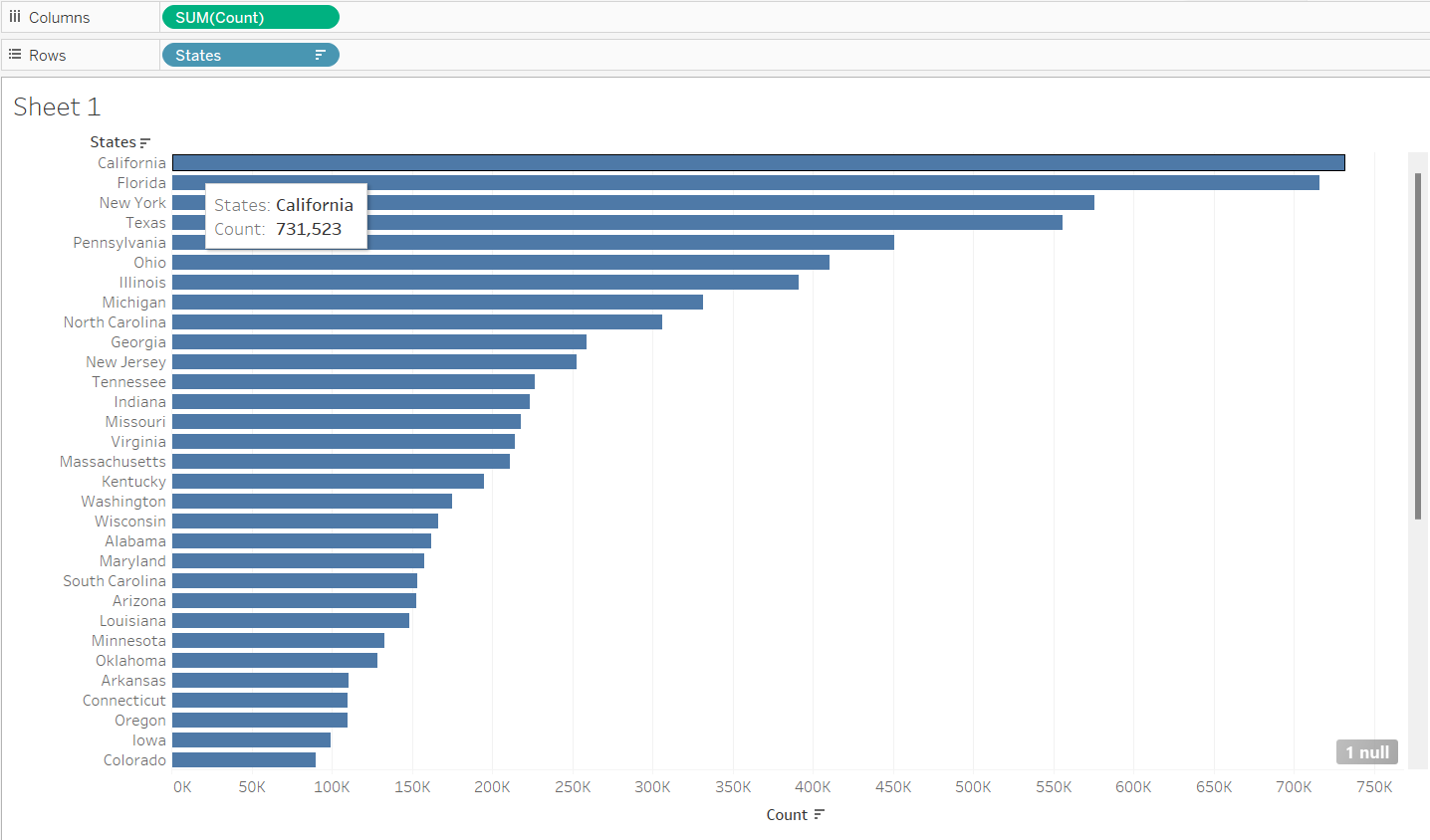
First, you need to connect to your data source, which is where your data is stored. Then, you need to create a map that shows the geographic area you want to analyze, such as a city or a region. Once you have your map, you can add your data to it, such as the locations of schools, hospitals, or crimes.

Next, you can create a density map, which is a way to show the distribution of your data across the geographic area. This is done by using density marks, which show the concentration of your data in different areas of the map. You can customize the density map by adjusting the color, size, and opacity of the density marks, as well as adding labels, legends, and other visual elements.

Finally, you can analyze the density map to gain insights into your data. This can involve using filters, sorting, and other analysis tools to explore the data in more detail. By following these steps, you can create informative, accurate, and impactful analyses that help you make better decisions based on your data.

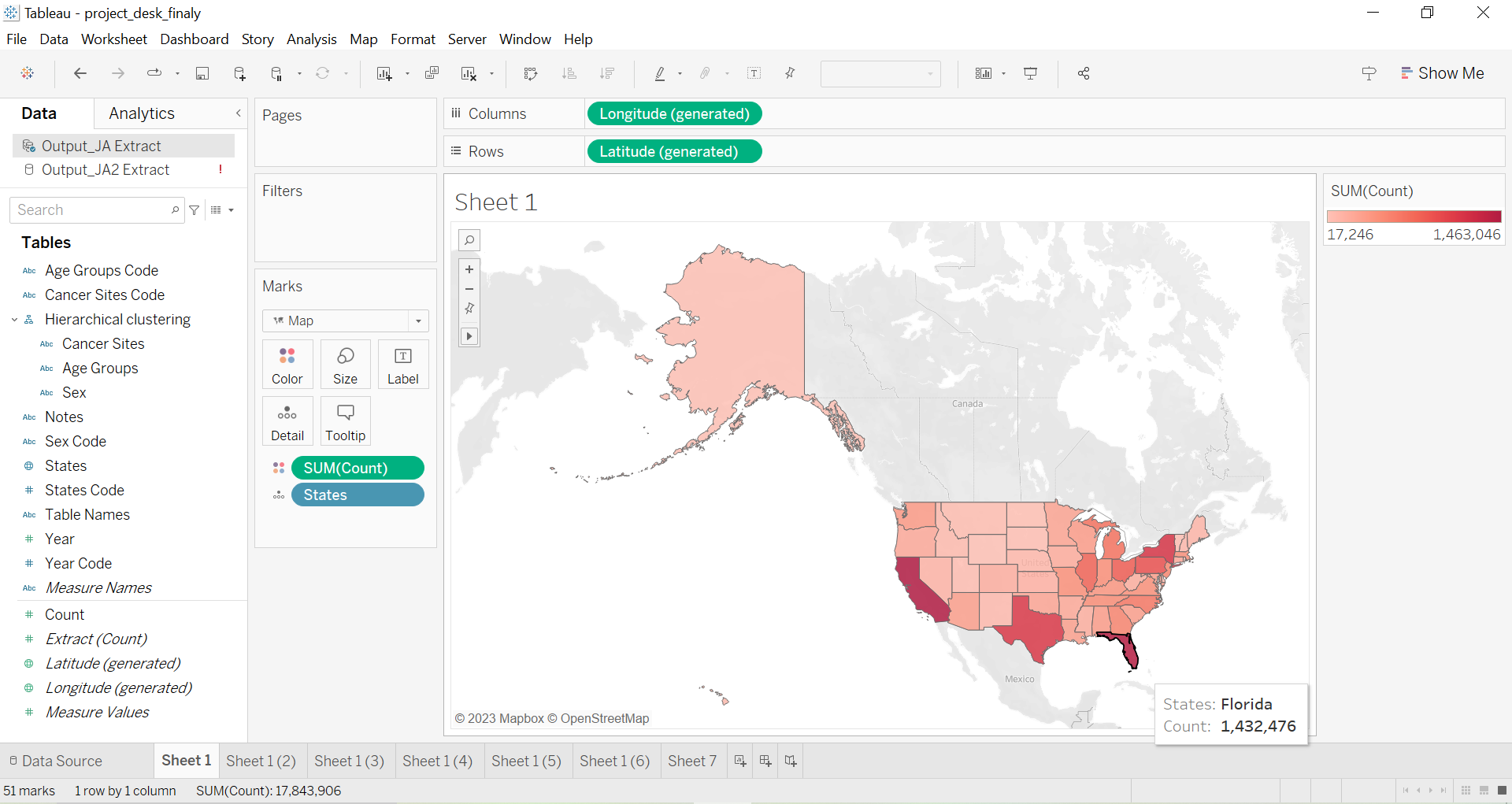
**Chapter 6: Horizontal Bars / Number of Cases per state**

In Tableau we use horizontal bar chart technique to visualize the number of lung cancer cases for different states. Here we easily visualize as California has the highest number of patients related to lung cancer over the years. We defined count from Measure field as column and States from Dimension field as rows. Then we sort data to visualize the state has highest number of patients related to Lung cancer over year.

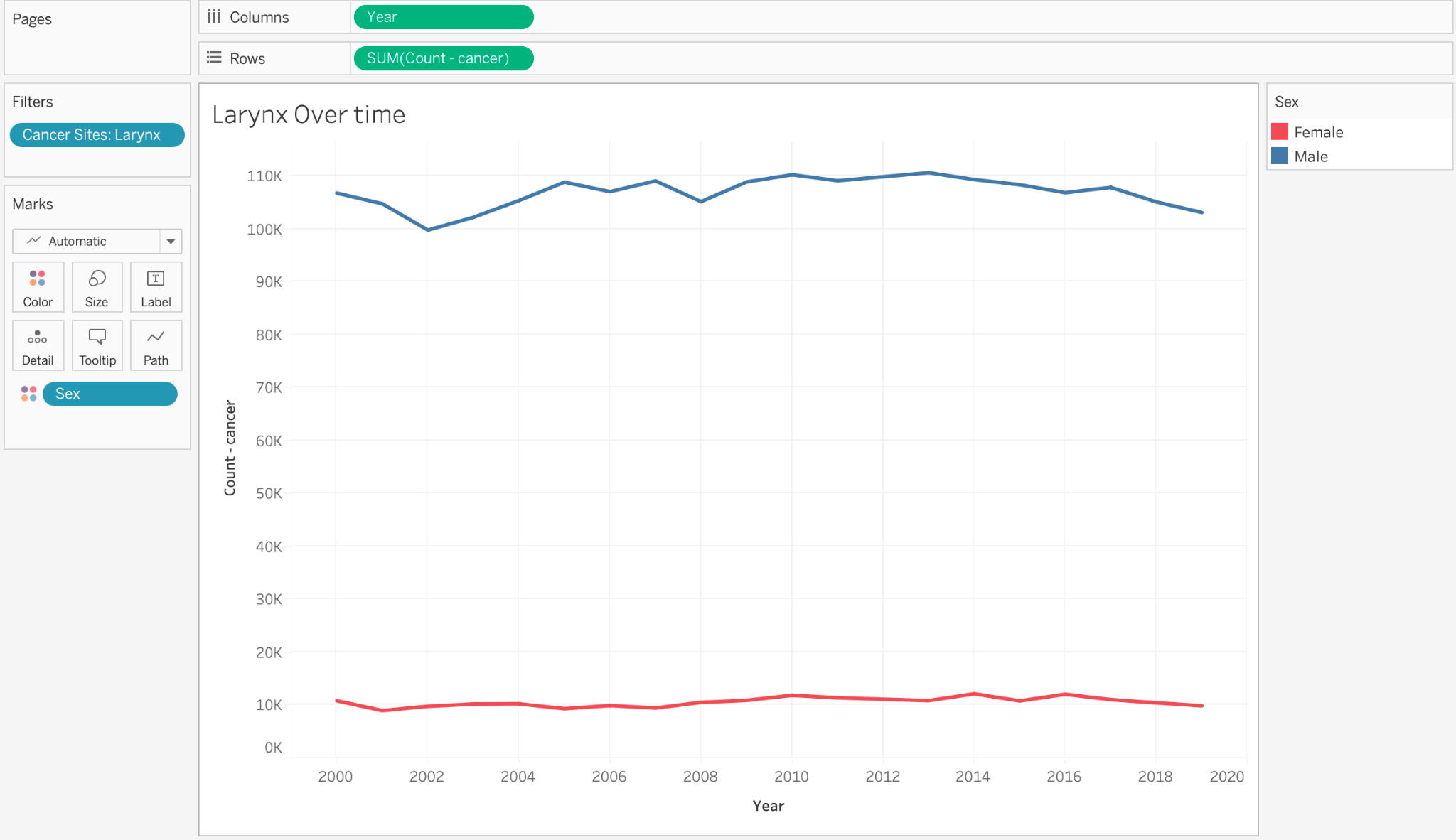
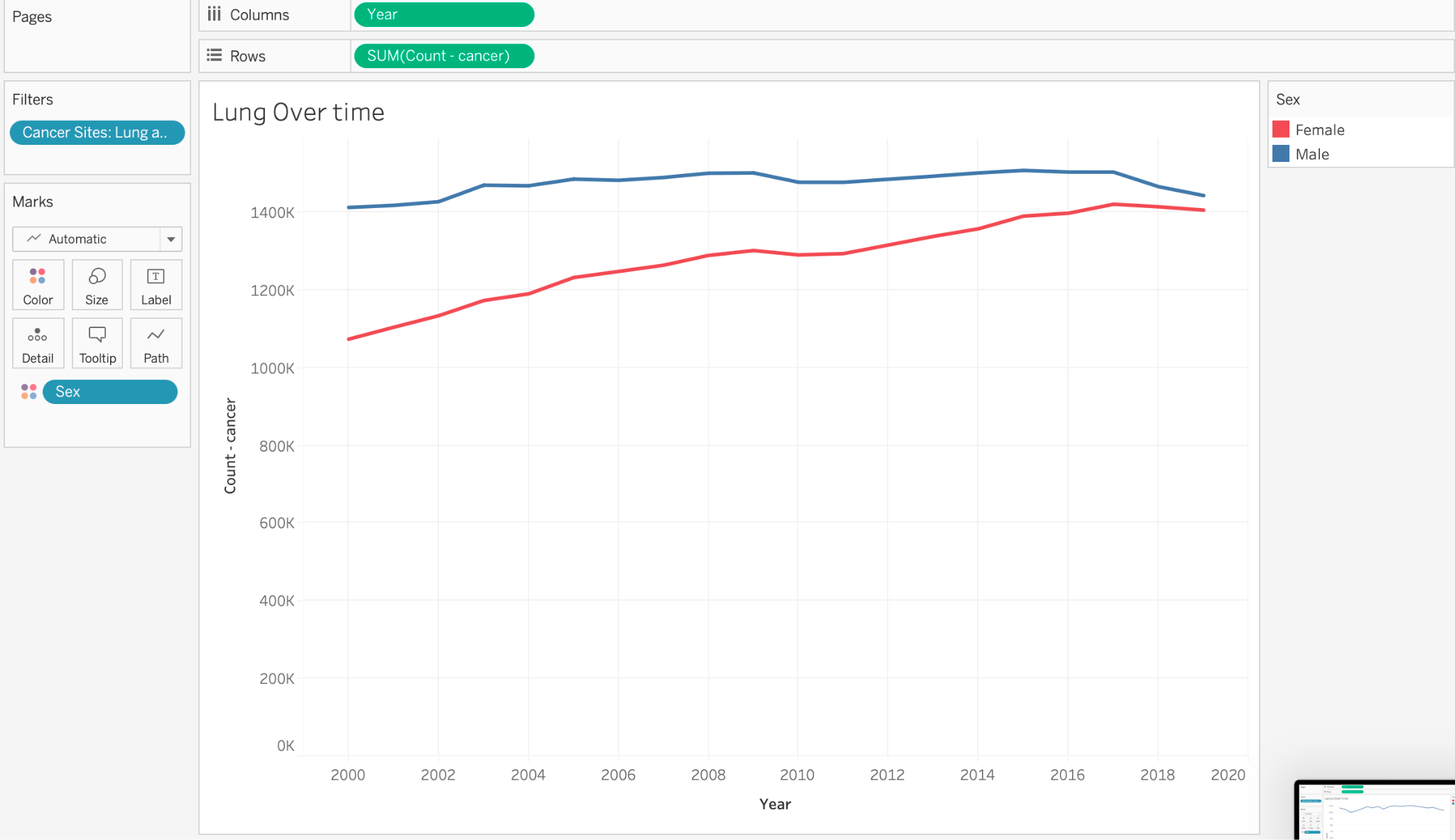
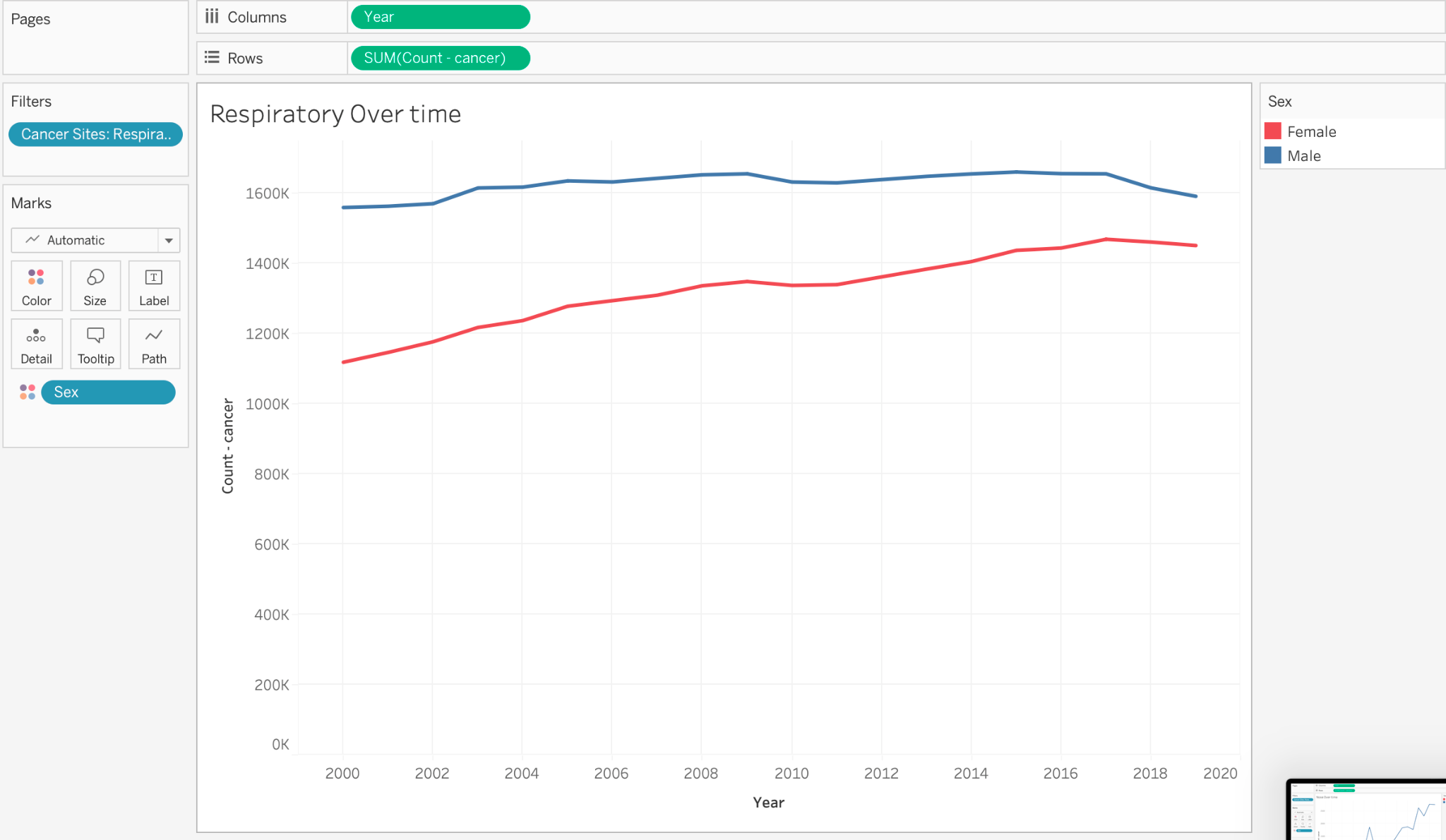


**Chapter 6: Maps Visualization**

Using maps can help us visualize how geographical location correlates with trends in data when geographic information is associated with the data. Maps can show us that Florida is one of the states with the highest numbers of cases, providing evidence that the outlier we found is genuine. This visualization prompted us to ask questions about why states like Florida, Texas, and California have high numbers of cases. Through research, we discovered that over 40% of adults in Florida use tobacco, while in California, there are issues with cannabis use, as evidenced by studies on emergency department visits and hospitalizations. For instance, in one study on drug use among drivers in fatal motor vehicle crashes in California, over 50% of drivers tested positive for cannabis.

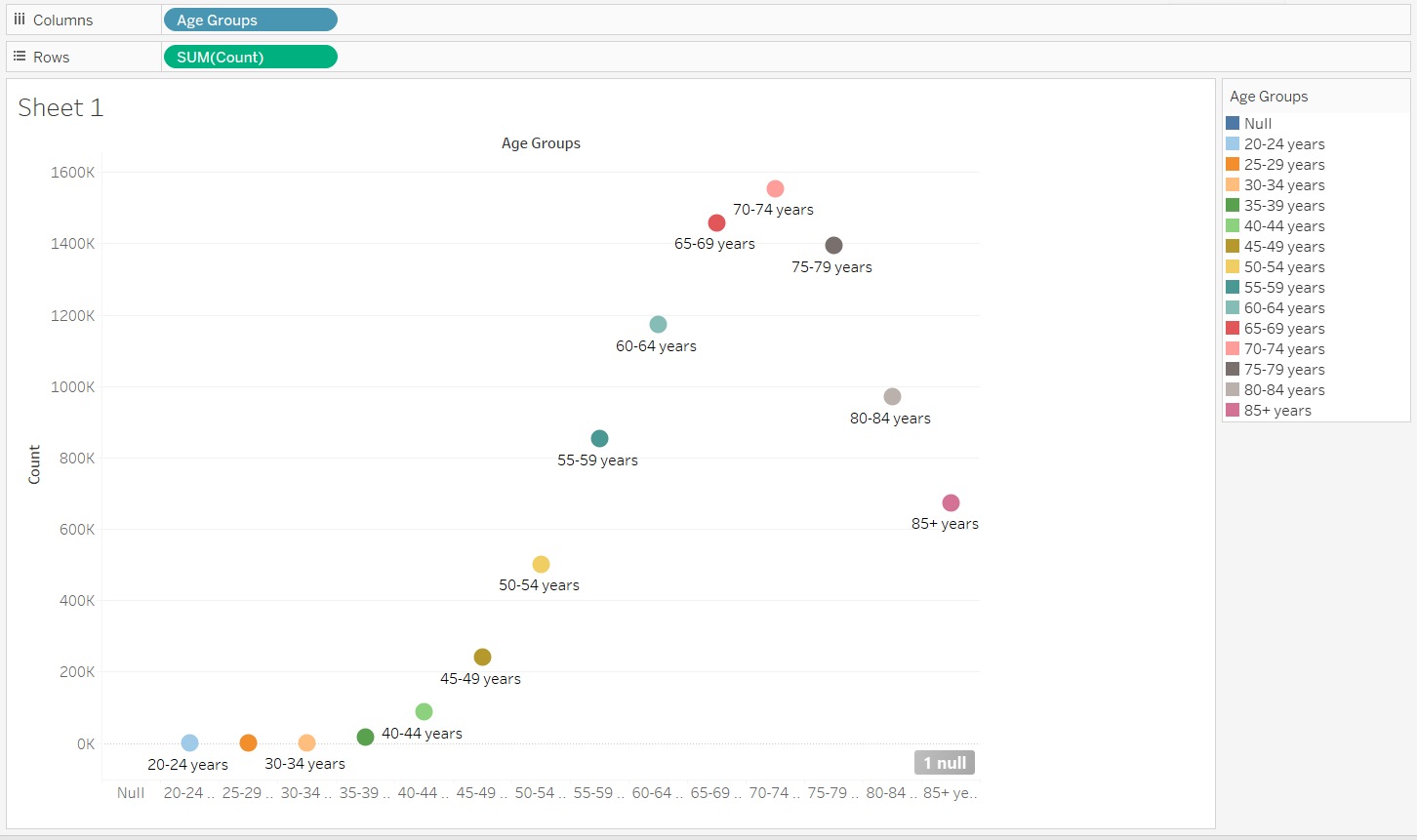


**Chapter 6: Line Graph**

To properly interpret the separate cancer types and their rates over time by sex, we selected to use a line plot. In tableau we created four sheets, one for each identified type of respiratory cancer. Lung, Respiratory, Nose, and Larynx. We filtered the data by those four types on each respective sheet. Then we selected the year as the columns and the count of the specific cancer type as the rows. We chose sex for the color and then selected the layered line plot to display this. This gave us four separate line plots for each cancer type with two lines for each sex. These were the following plots. 

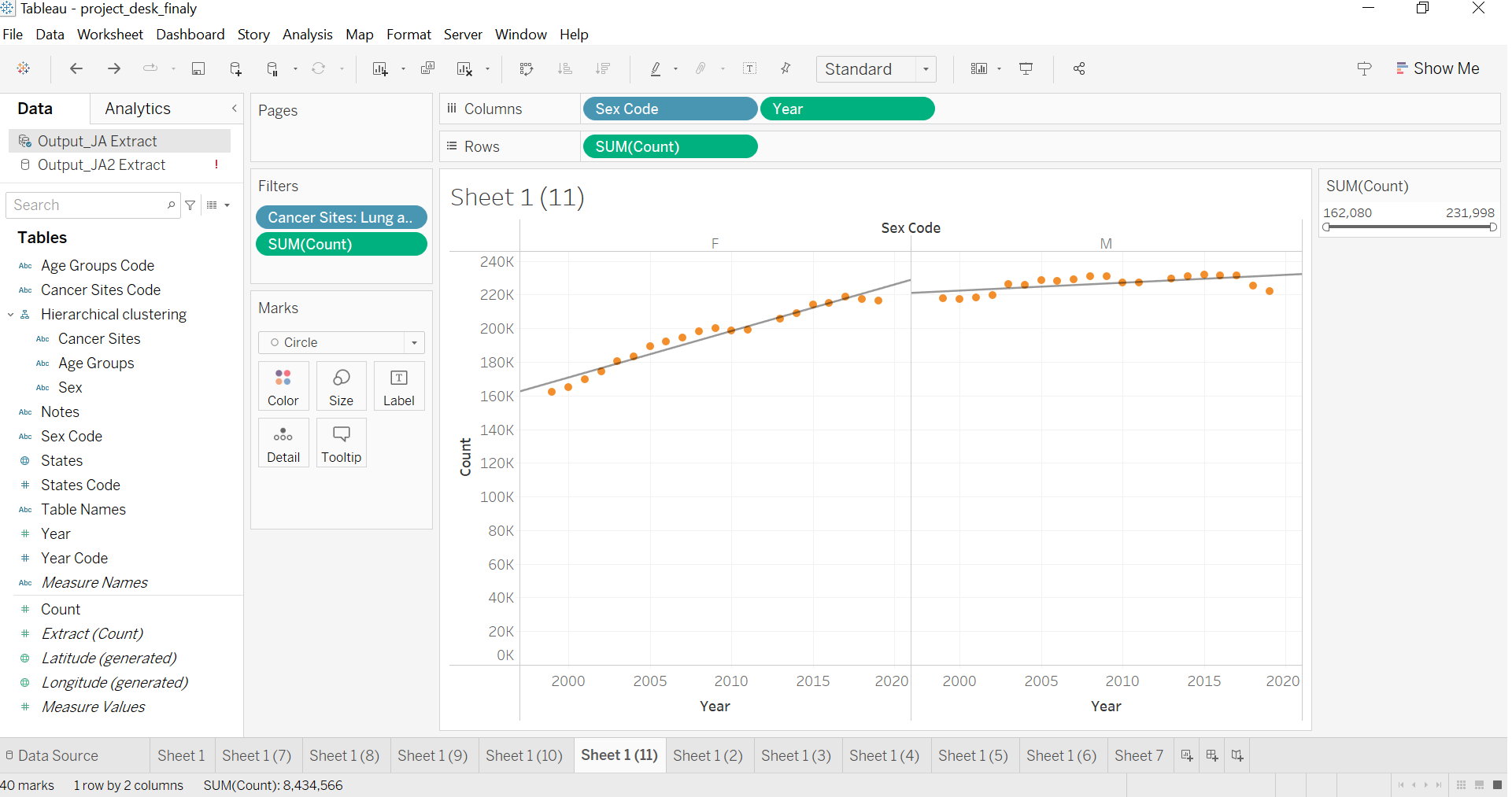
**Chapter 7: Data embedding/ Number of cancer patient by age group**

In tableau we use Data embedding with circle mark technique to visualize which age groups have the highest number of patients related to lung cancer. Here we easily visualize as the age band 70-74 has the highest number of patients related to lung cancer. We defined count from the Measure field as rows and Age Groups from the Dimension field as columns. Then we assigned color to the Age groups field to easily differentiate different Age groups data.



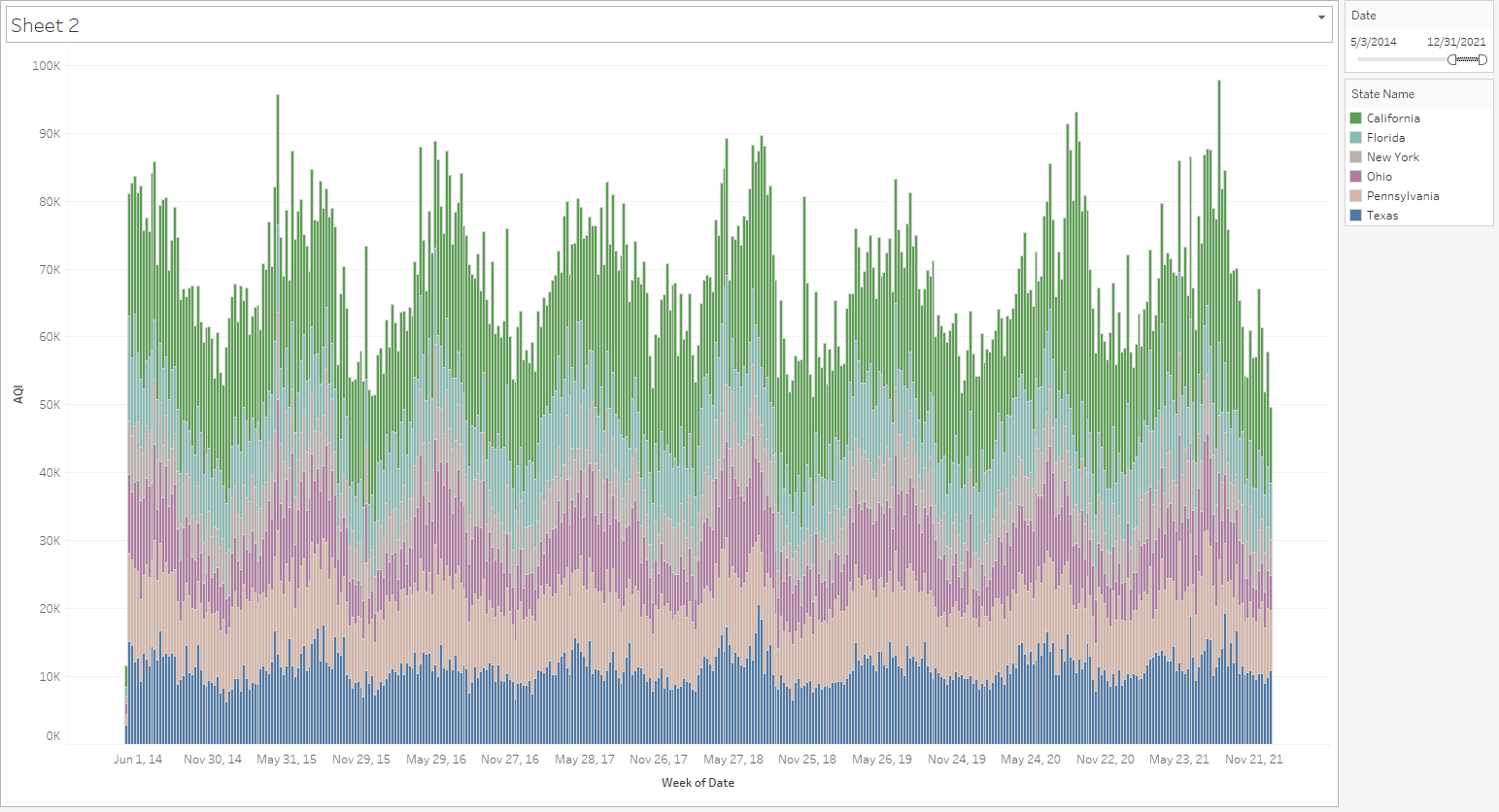
**Chapter 7: Correlation technique**

Analyzing the relationship between the data in our project is critical to provide us with more information and understand any correlation between the variables. To explore this, we used correlation techniques between years and the total count number and compared the results for both genders to see if they had the same correlation. Our analysis revealed a strong positive correlation between years and the number of cases for females, while males showed only a weak correlation.



1. Chapter 9 – Aggregation –

In Tableau we use Aggregation techniques with Maps and pi-chart to visualize female vs male ratio of lung cancer related patients with respect to different states. First, we filter sex and state data from the Dimension field to remove any null values. After selecting the pi-chart we mark states as detail and sex as color to easily visualize male vs female ratio of Lung cancer related patients at different states.

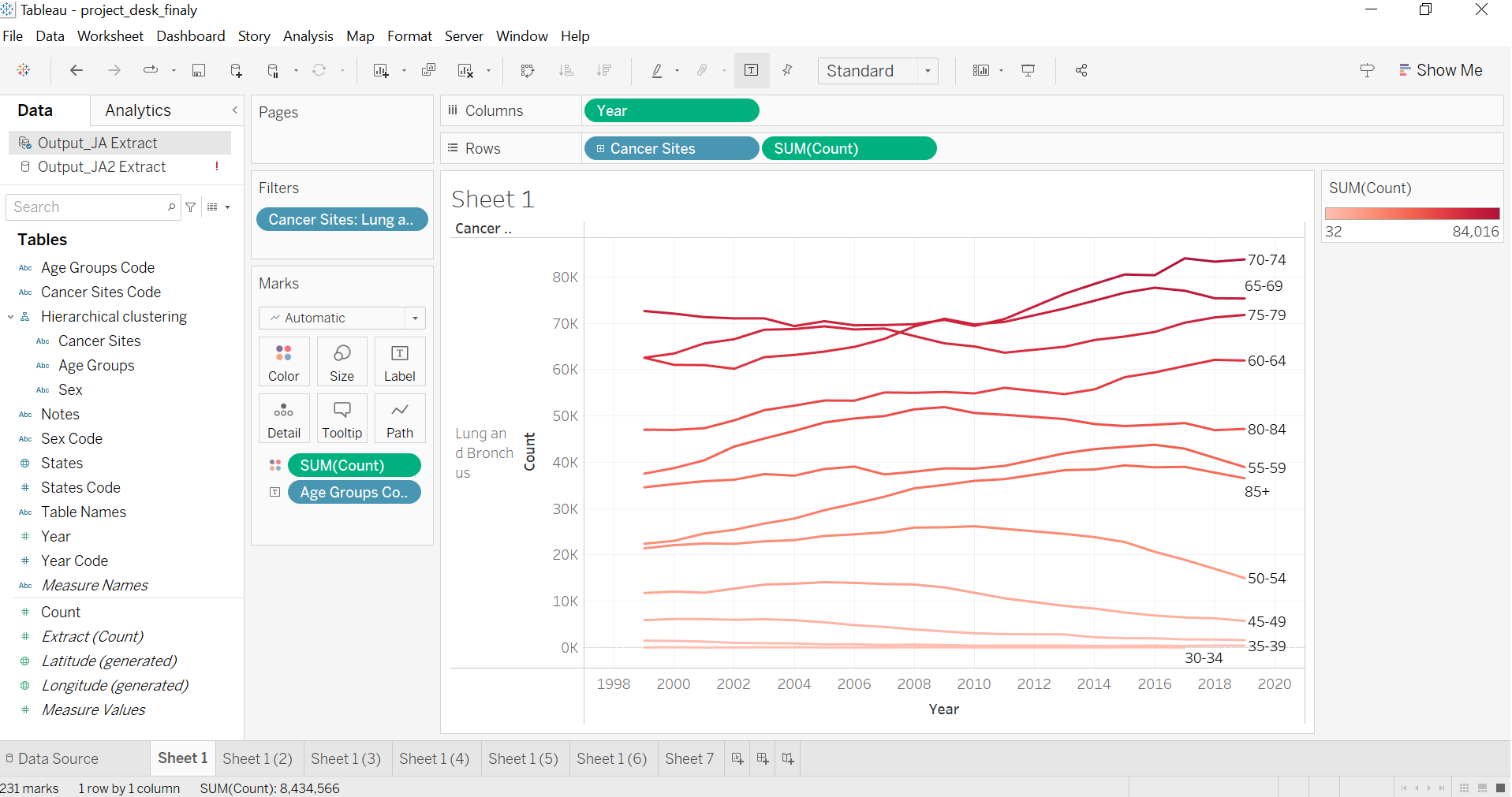
**Chapter 7: Multiple Frequency Histogram**

Here we can see a multiple frequency histogram of the highest AQI pollution states. We can see that California leads every other state by a very far margin, and then followed by Florida, New York, Ohio, Pennsylvania, and Texas. The size of the bar indicates the AQI value.



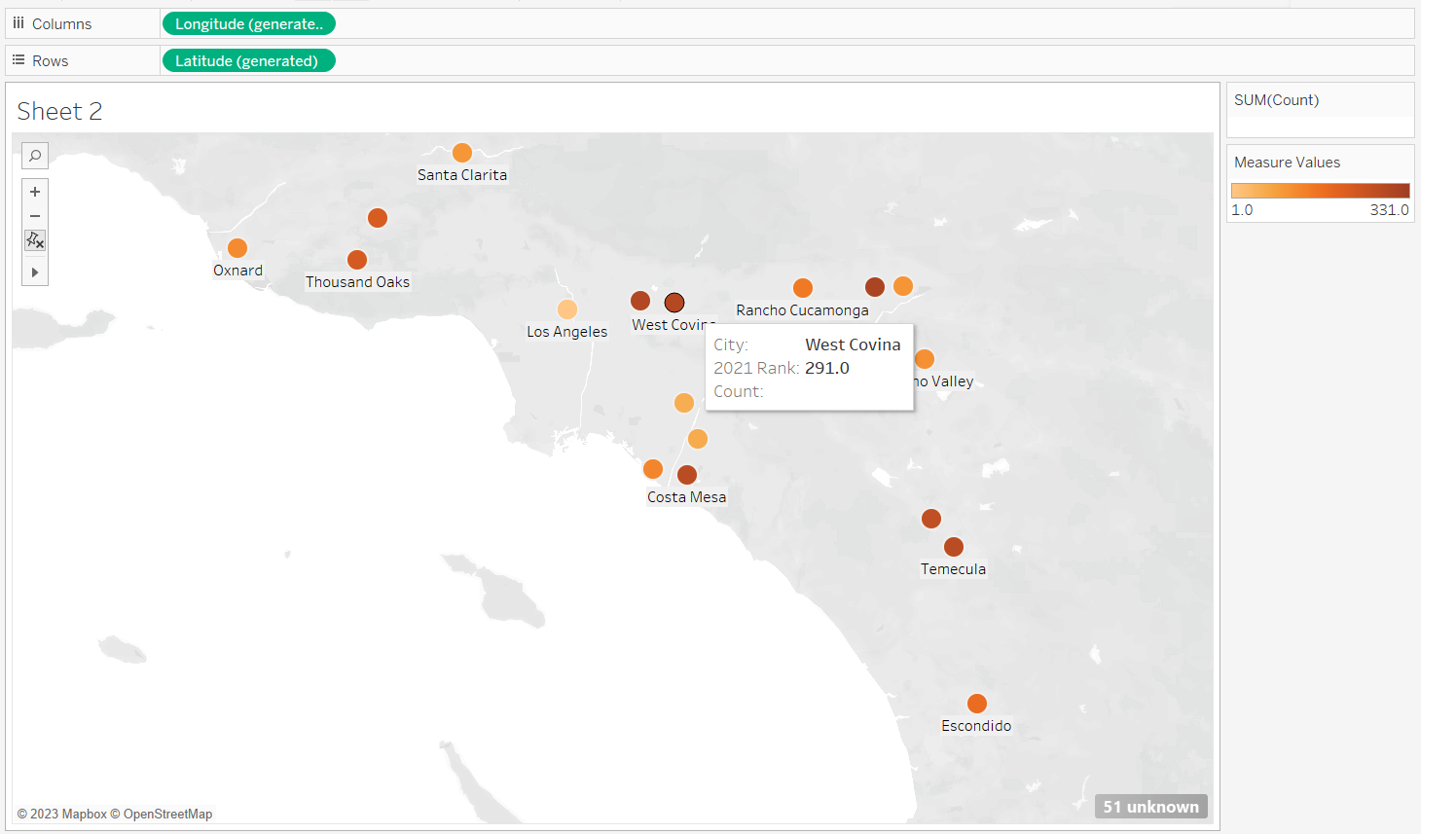
**Chapter 9: Time plot or time graph**

In our project, it is important to use a time plot or time graph to get a clearer picture of the situation from 1999 to 2019 and understand which age group had the highest number of cases in the past 20 years. This graph provided us with a clear answer that the age group between 70-74 had the highest number of cases.

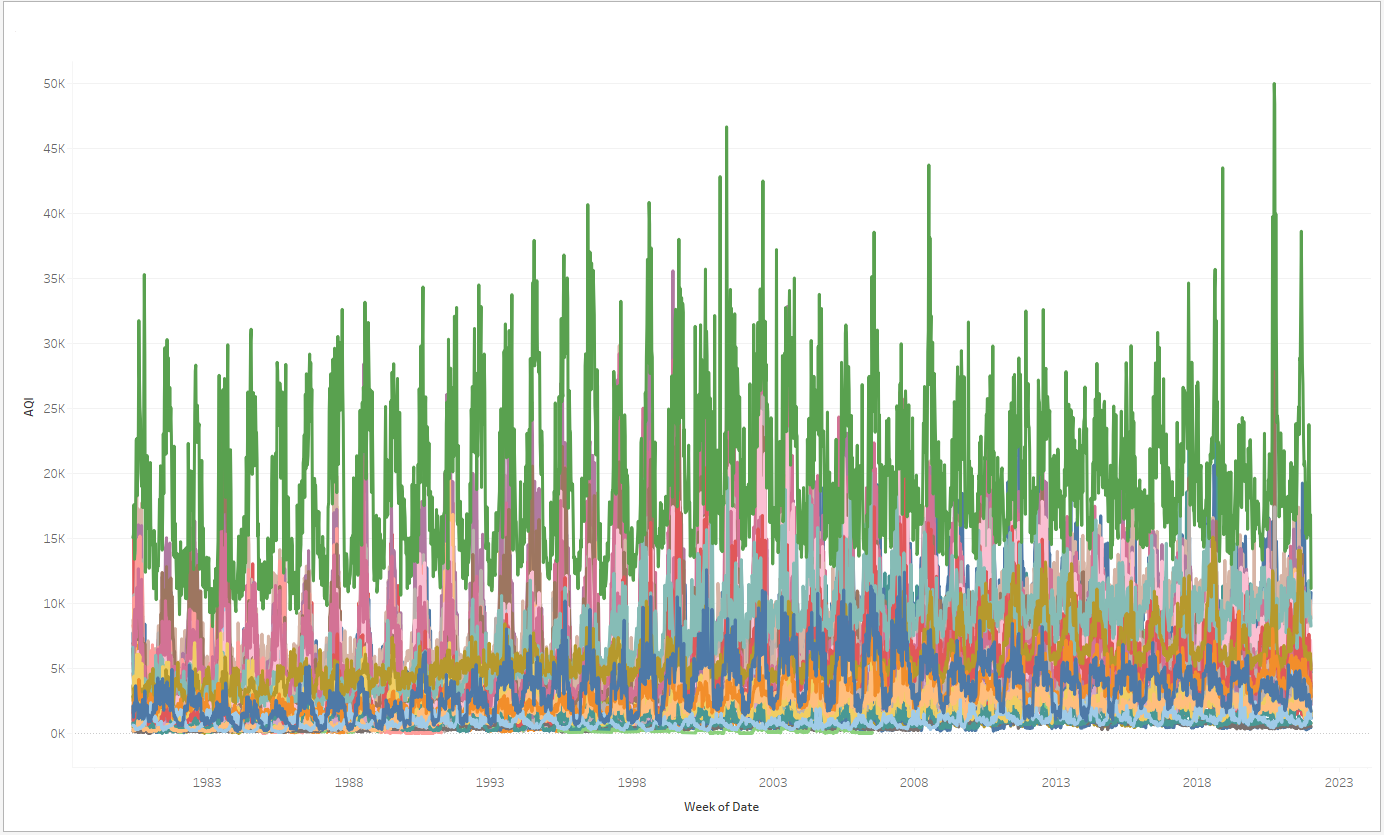


**Chapter 10: Gridding**

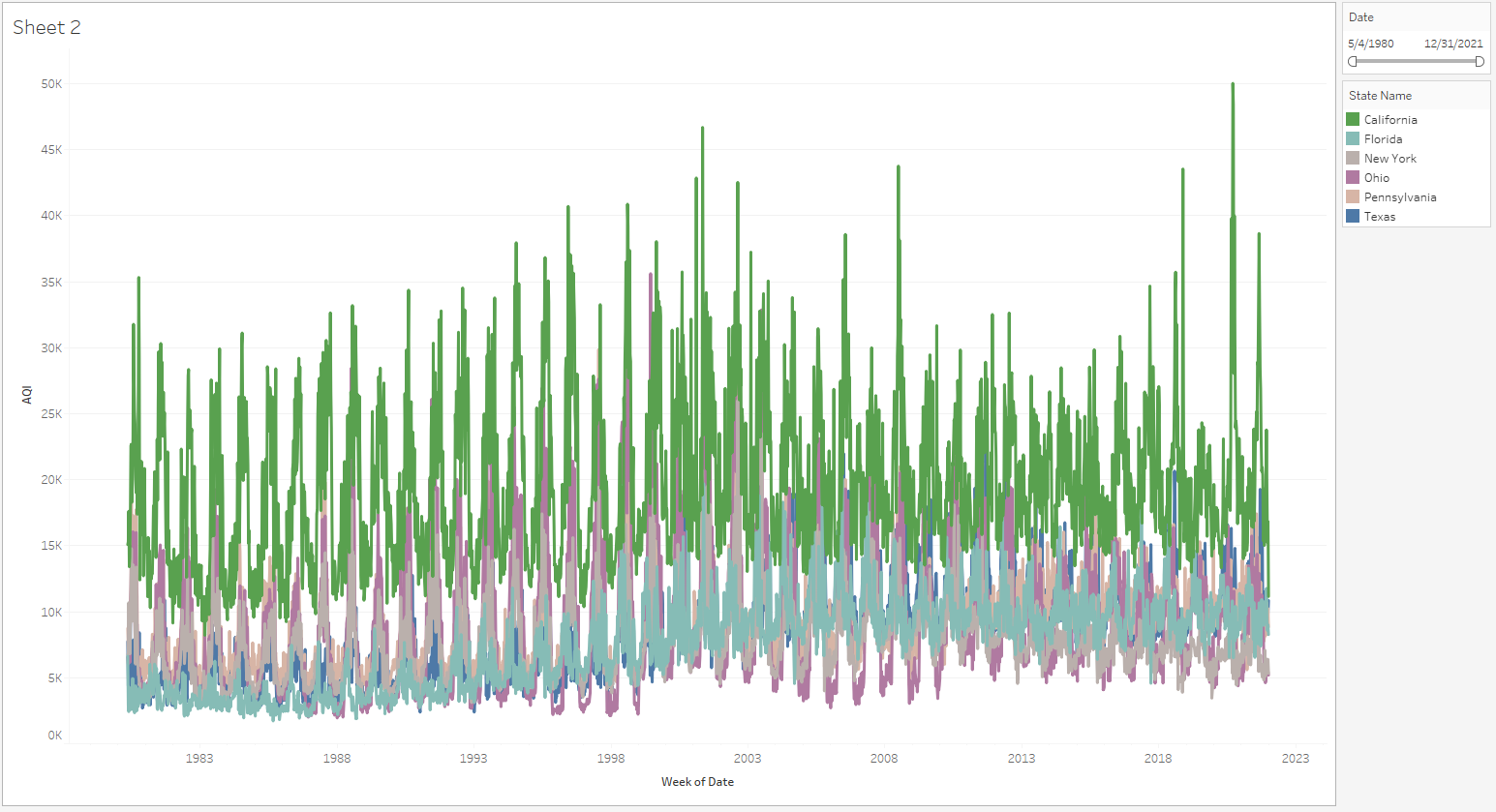
In Tableau we use Gridding technique with Maps and circles to visualize the number of cancer patients in different regions of California state. First, we filtered state data from the Dimension field to select only California. Then we connect USA city data at the data source tab. Then we mark count data from the Measure field to color and mark city field to label. Here with different shades of color we can easily visualize which region of California state has a high count of the number of patients related to Lung cancer compared to other regions. Dark color shade represents high count of Lung cancer related patients` and light color share represents low count of Lung cancer related patients.



**Chapter 9: Temporal distribution**



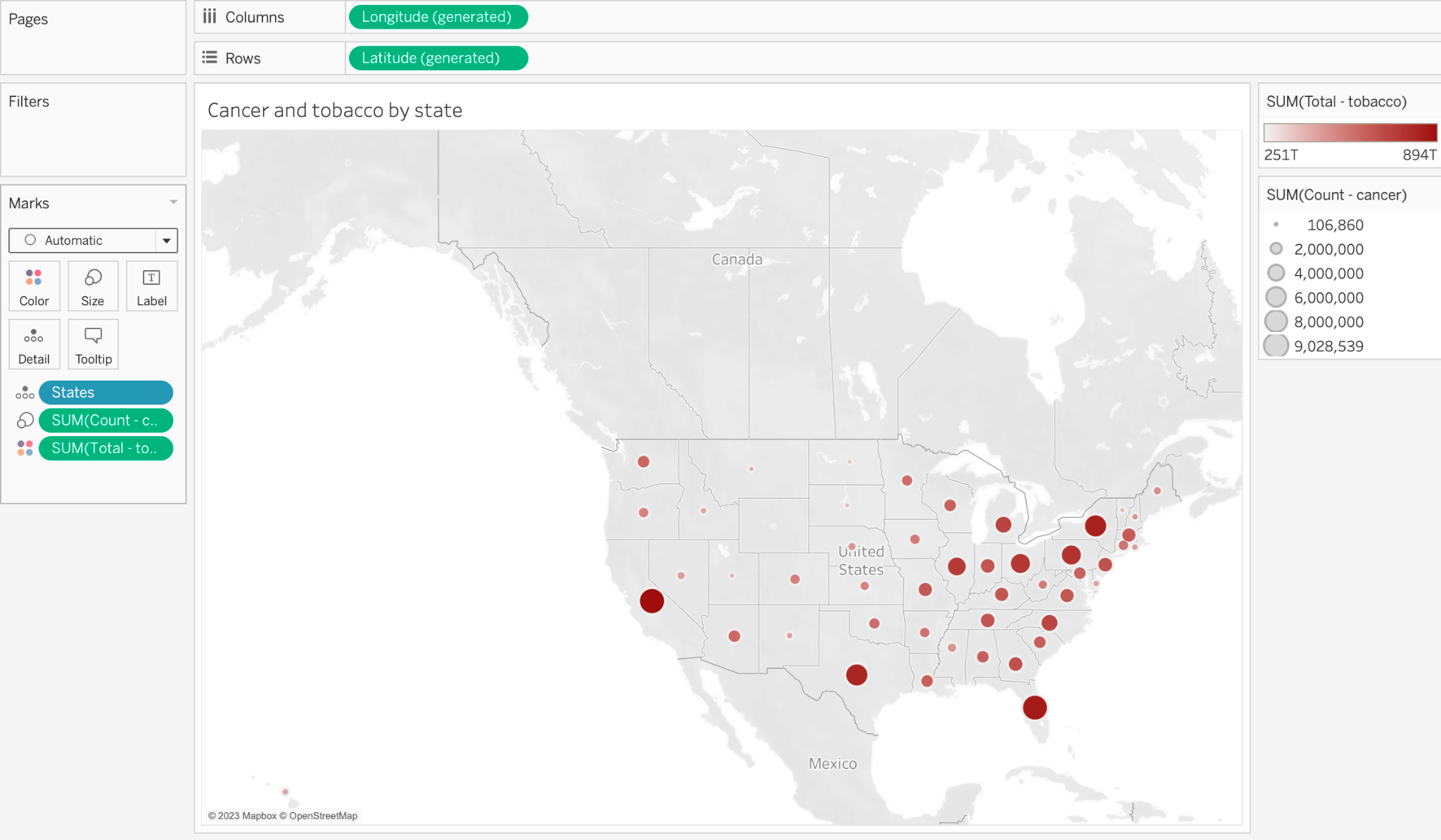
Here we can see a pattern emerge using a temporal filtering technique, where the data is averaged per week. We can see a frequency of Air Quality Index data that displays two things, one is which states have the highest fluctuation in AQI amplitude, where the peaks mean the air quality is very poor, the higher the the scale goes. Observe the Legend for the filtered image above below:



The states with the poorest air quality index are California, New York, Ohio, Texas and Florida. We can see a pattern where the air quality index is worst during summers and improves drastically in the winter. Since AQI is determined using particulate matter (PM) 10, PM2. 5, Ground level Ozone (O3), Sulphur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), lead (Pb) and ammonia (NH3), we can see that during hot seasons these gasses and particulate matters increase in the air due to the heat introduced to the environment.

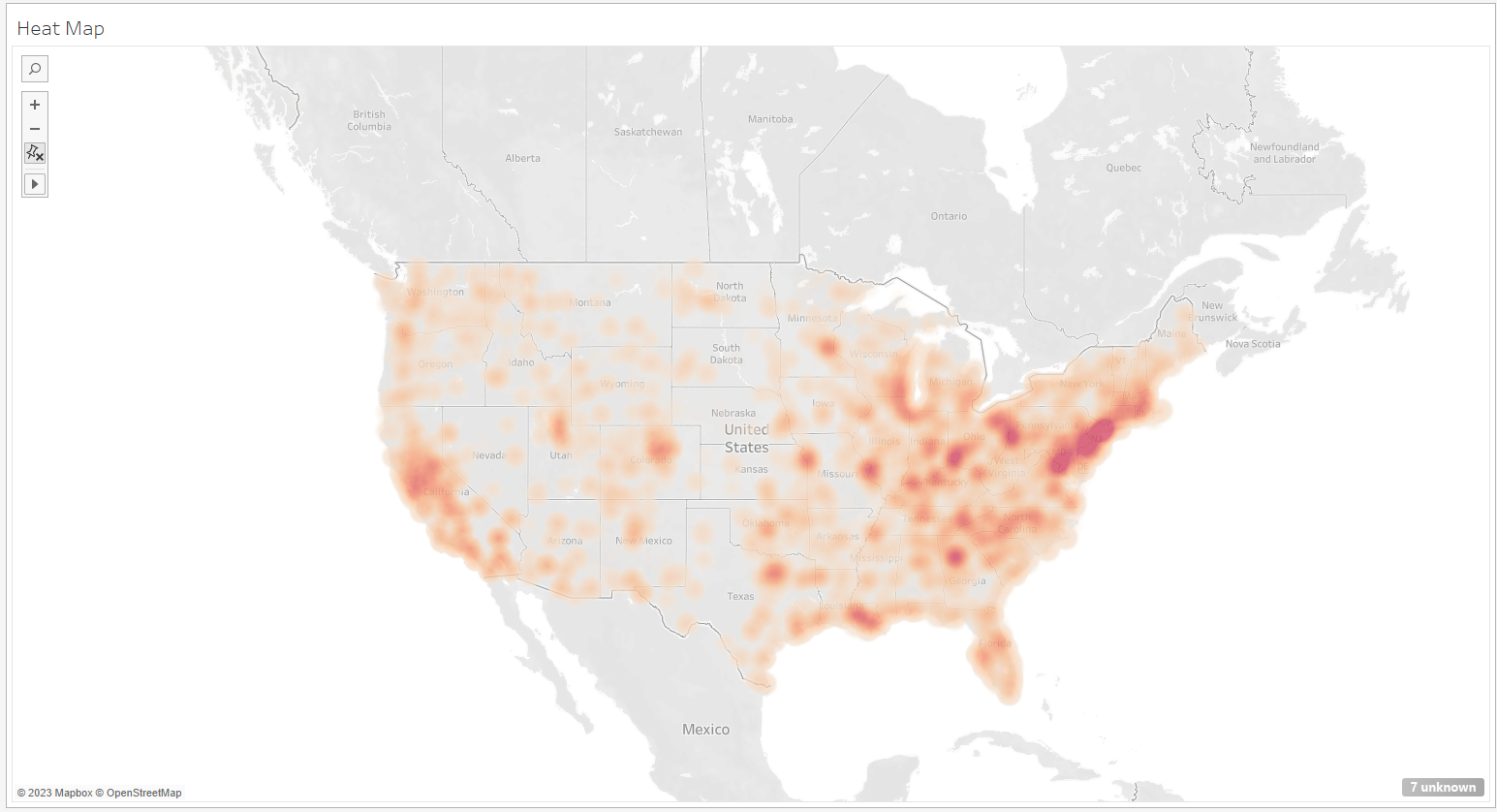
**Chapter 10: Spatial Joining**

To combine the tobacco rates and the cancer rates by state, we are implementing spatial joining. This will allow us to view both sets of data in one shared space and to draw conclusions on how the two are related geographically. To do this in Tableau we selected, state for a geographic field, and then we selected both the count for cancer instances and the total for tobacco use. This allowed us to select the map visualization and then have the following result.



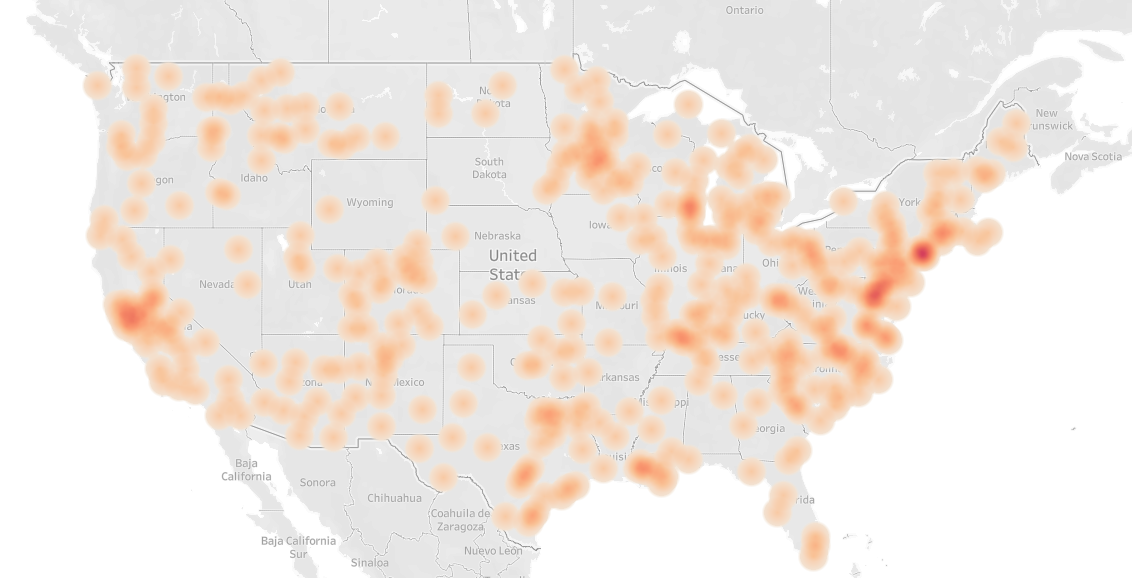
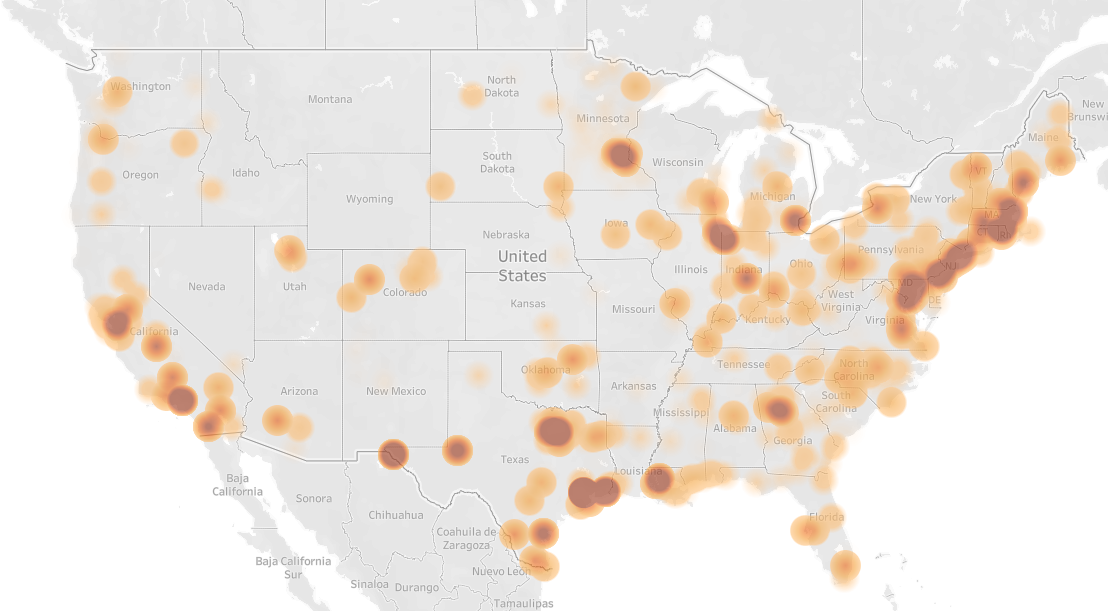
This shows the Cancer rates as the size of the circle and the tobacco rate as the color of the circles. Therefore, the darker and larger circles have the higher rates of both cancer and tobacco.

**Chapter 10: Rastering**



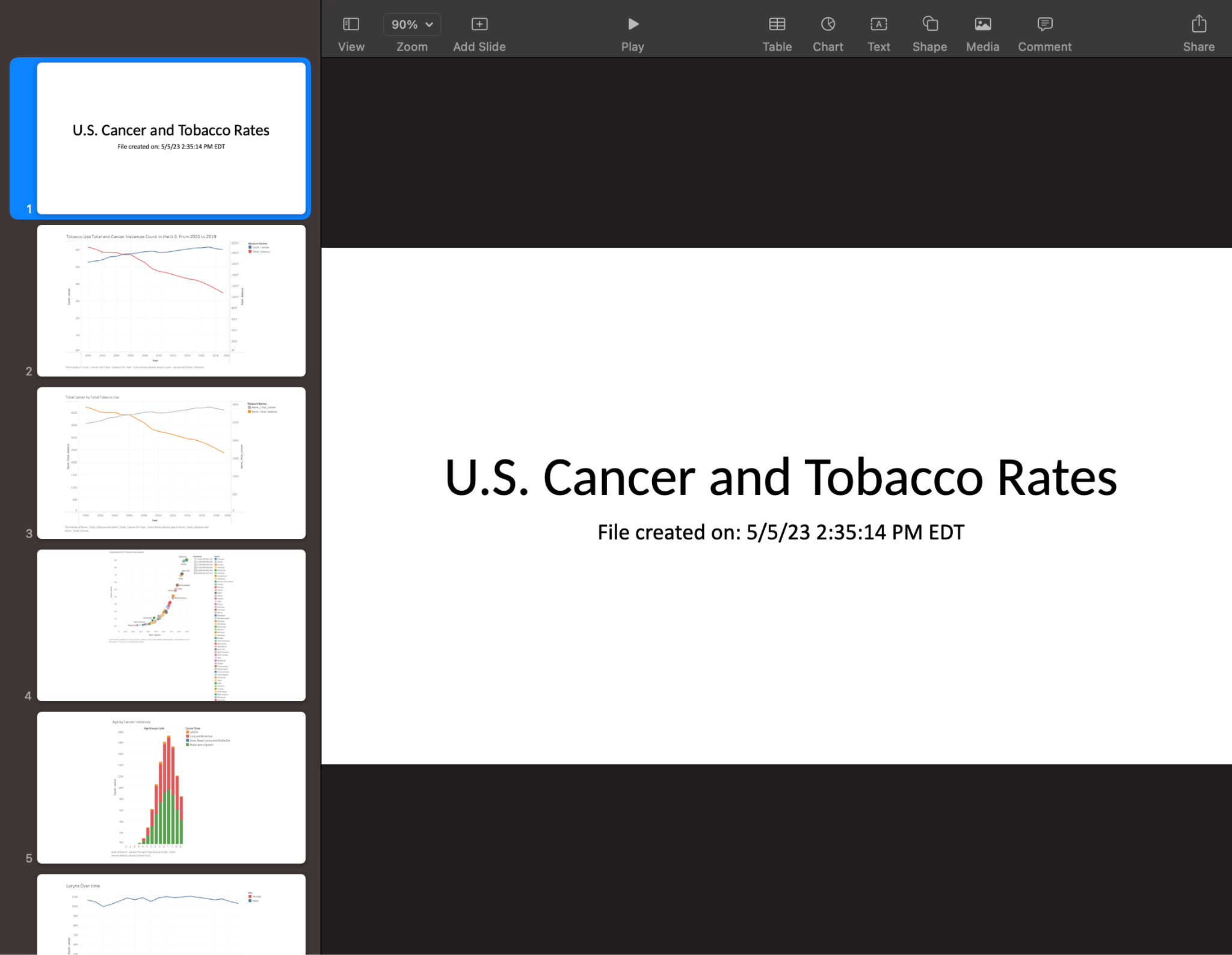
Here we can see an elevation map or heat map of the highest AQI in red. Again, the higher the AQI the more hazardous the air is. We can see quite clearly that the New England area, California and the border southern states seem to have the worst AQI. This indicates that the air quality is highly related to population density. If you notice here though, Florida is not a hazardous state, but still has a high rate of lung cancer. This is likely due to the age of the people living in the state.

We can also see below the VOC and HAP content (respectively):



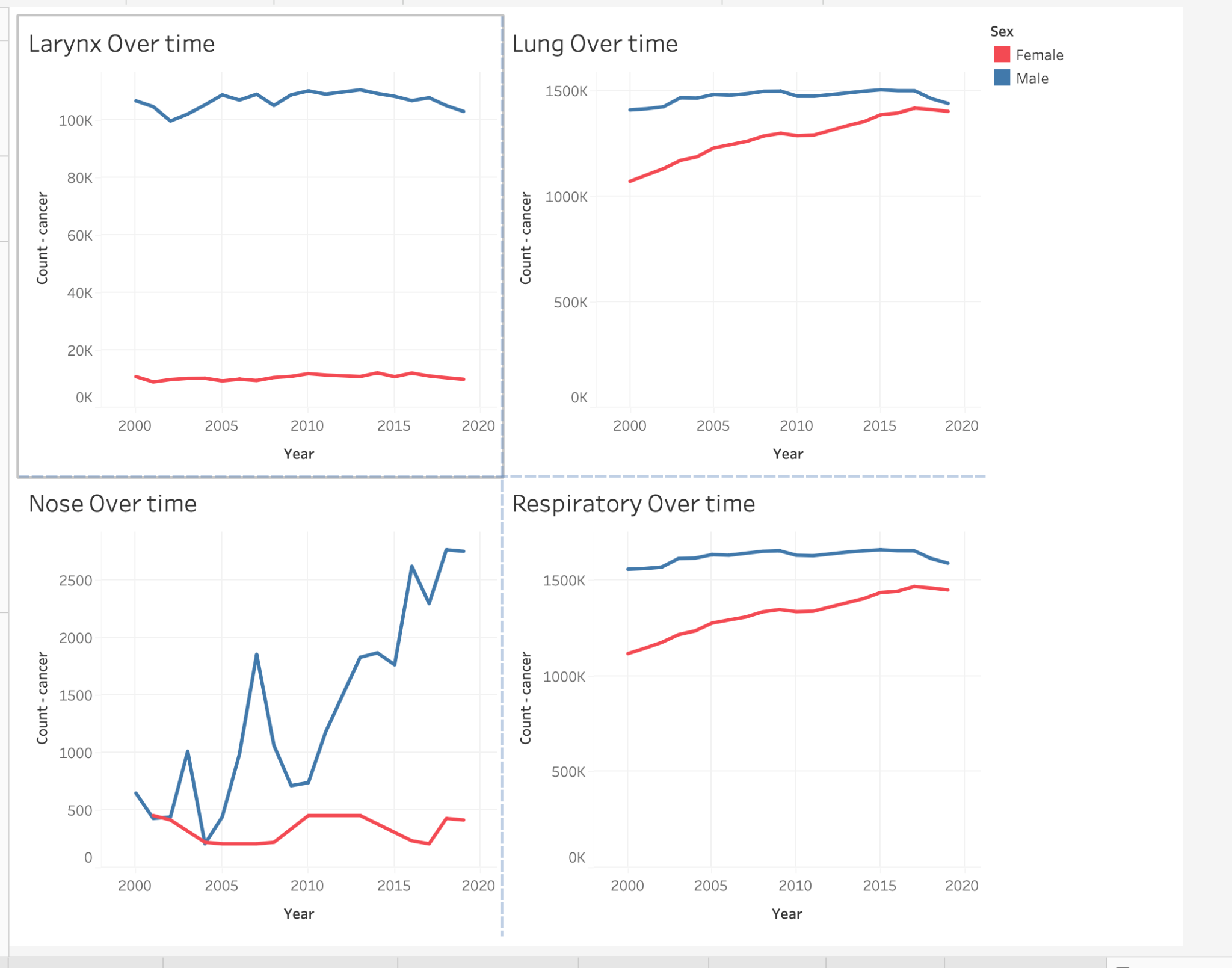
**Chapter 11: Slideshow**

In order to tell the full story of the cancer rates and the tobacco rates in the US over time, we exported all of our dashboards and sheets to a powerpoint. This allows each visualization to be viewed in the context that us the creators wished, but leaves the interpretation up to the viewer.



**Chapter 11: Static Storytelling**

Another method used for storytelling is static storytelling. We chose to demonstrate how the different cancer types have impacted the different sexes over time. Using our line plot from earlier, we are able to create a dashboard which can display all line plots simultaneously. This allows the viewer to simultaneously see all types and compare the rates at which they have affected males and females over time.



**Chapter 11: Animated Storytelling**

[Here](https://drive.google.com/file/d/1PXLqpp2zT18S53jFkuKKtMlsuk7S_qN3/view?usp=sharing) is a link to this Animation of AQI data over time for the past 42 years:

This animation shows the progression of AQI data over the past 42 years and we can see that it has been increasing slightly, but overall, stayed the same.

| Lecture Number | Task completed | Description | Who it was completed by |
| --- | --- | --- | --- |
| 2 | Filtering | Removed similar field State Code, Year Code, Age Groups, Sex Code | SOHIL |
| 2 | Filtering by fix columns | change some of the columns' names to do join operations with others files | Diaa |
| 2 | Explore the data to identify patterns | Comparing counts by age range to identify ages with higher rates of cancers | Janna |
| 3 | Investigating data to find outliers | find outlier | Ahmad |
| 3 | Sort | Sort count field to see number of cases/patient reported for respective age band. | SOHIL |
| 3 | Join | join US Cancer instances with tobacco use by year | Janna |
| 4 | Normalization/Scaling | to compare tobacco rates with rates of cancer instances | Janna |
| 4 | Missing data | Check if there are missing data in the dataset | Diaa |
| 4 | Processing data | Check if the data are balanced/fit/clean the data | Diaa |
| 5 | sequence/graph | to check for the data if there any sequences and apply graphing | Diaa |
| 5 | Hierarchical clustering | Builds a hierarchy of data subsets | Ahmad |
| 5 | density base | Two data items are treated as neighbors/threshold | Diaa |
| 6 | using Maps | visualize any kind of location information | Ahmad |
| 6 | Horizontal Bars | visualize number of cases per state. California has highest number of cases | SOHIL |
| 6 | Line graph | Cancer Types over time by Sex | Janna |
| 7 | Data Embedding | Number of cancer patient by Age Group. Age band 70-74 has highest number of lung cancer patient | SOHIL |
| 7 | order relations\_Correlation | Attributes that are highly interrelated or independent from others | AHMAD |
| 7 | Multiple frequency histogram | area/high density region analysis using cancer data, air data, desaster, etc. | Ali |
| 9 | Time plot or time graph | analysis task is to understand how  they are distributed in time | Ahmad |
| 9 | Aggregation | Female vs male patient ratio of different states. Used pi chart to show ratio | SOHIL |
| 9 | Temporal distribution | Filtering data temporally and displaying it cyclically over a geomap | Ali |
| 10 | Gridding | Number of cases areawise. Showed different cities of california state for count of patients. | SOHIL |
| 10 | Spatial Joining | US cancer instances and tobacco use by state | Janna |
| 10 | Rastering | US air quality data based on air quality monitors PM10 and Ozone health | Ali |
| 11 | Static Storytelling | US cancer types over time by sex | Janna |
| 11 | Slideshow | Visualizations explaining relations between cancer and tobacco | Janna |
| 11 | Animated Storytelling | Air Quality over time | Ali |