**REDESIGN PROJECT**

**GROUP 1**

**KULDIP GADAPA**

**SAI TEJASWI BELLAPUKONDA**

**BHARGAV RAM NARA**

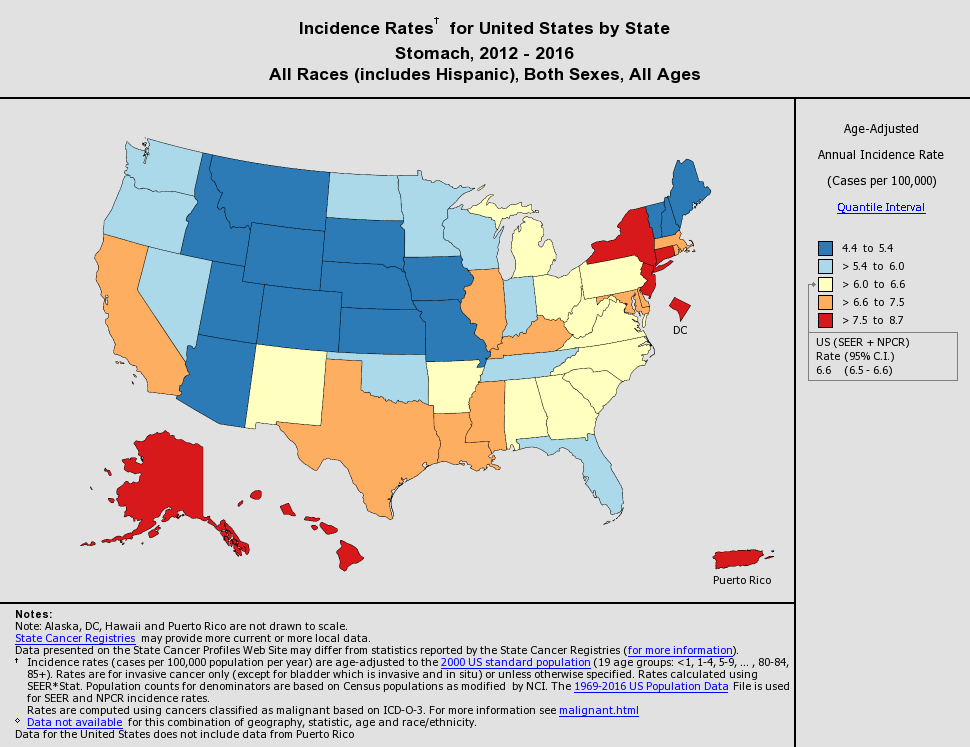
**Project: Redesign and Analysis of Incidence Rates on Stomach Cancer In USA**

**Introduction:**

Data Visualization is a graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Our eyes are drawn to colors and patterns. Our culture is visual, including everything from art and advertisements. Data Visualization is another form of visual art that grabs our interest and keeps our eyes on message. When we see a chart, we quickly see trends ang outliers. If you have ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization.

The redesign project is about analysis of incidence rates on stomach cancer in USA. The data is collected from CDC (State Cancer Profile Website) and the below maps are showing the analysis of stomach cancer incidence rates in USA. The graph designed in website is not clearly visualized which leads to misconception of data.

**THE BAD GRAPH:**



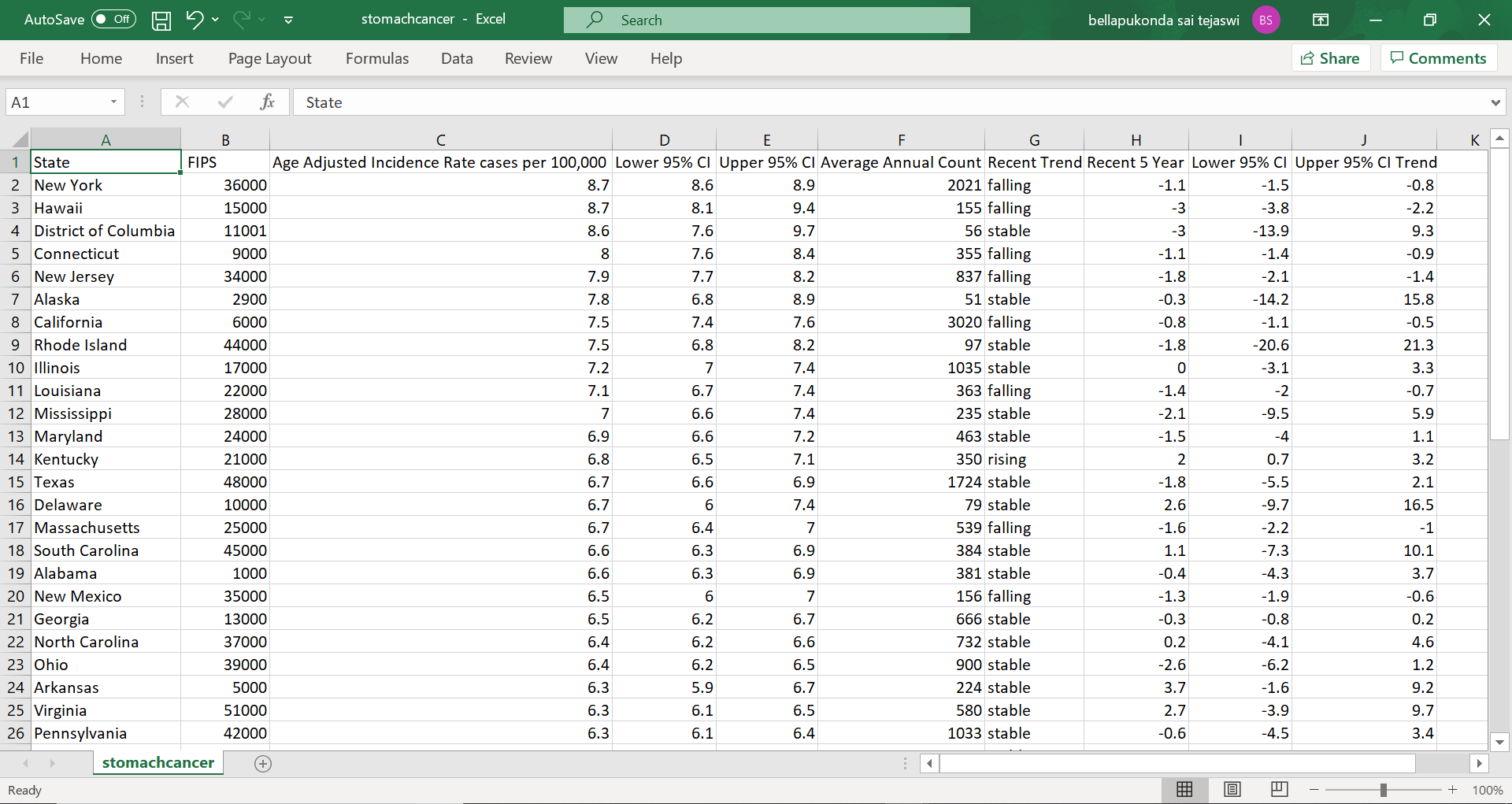
**FIGURE 1: BAD GRAPH**

The above bad graph is taken from State cancer profiles website. The graph is about stomach cancer incidence rate for each state and for all races, gender and ages in USA from 2012 to2016. This report objective is from Healthy People 2020 project done by the centers for disease control and prevention. This project is processed in cancer research, detection and treatment have resulted in decrease in both incident and death rates for all cancers. The cancer objectives suggest the importance of promoting the evidence-based screening for stomach, breast and cervical cancer by measuring the use of screening tests identified in the US Preventive Task Force (USPSTF) recommendations. The data set studied is about the incidence rates for stomach cancer in USA from 2012 to 2016. The graph is not an appropriate graph to represent the data.

**Drawbacks in Bad Graph:**

The main drawback is less information is delivered from graph. The graph does not show the actual values, the range and there is no proper labelling too. Hence conclusions cant be drawn.

**DATASET:**



**FIGURE 2: STOMACH CANCER DATASET**

The required visualization is made from the above dataset which is obtained from CDC Database where the data is collected from 1969 to 2019. The dataset includes data like the names of the states in USA, Age adjusted incidence rate cases per 100,000, Recent Trend, Average annual count, Confidence intervals for age adjusted incidence rates, Recent 5-year trend in confidence intervals and incidence rates. This dataset is a CSV file.

**REDESIGNED GRAPHS:**

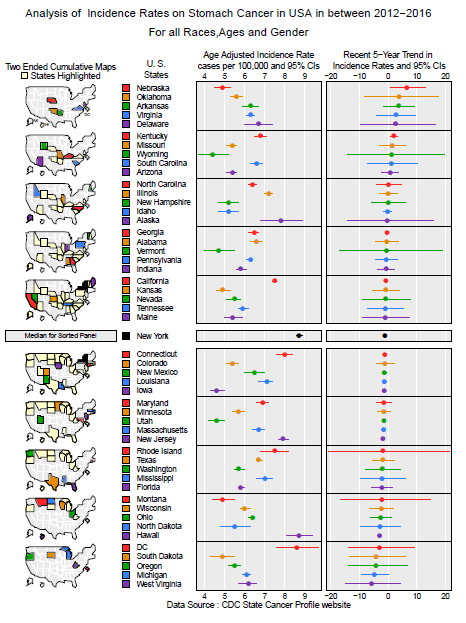
SOFTWARE AND PACKAGES:

For redesigning the bad graph, we have used R studio as software. In R Studio, we installed and loaded packages like micromapST and tidyverse in order to produce visualizations. Firstly, we cleaned the dataset and read into R Studio as data frame and then by implementing necessary code we get the required visualizations.

VISUALIZATION USING LINKED MICROMAP:

The linked micromap graphics package provides an easy and quick means of creating linked micromaps for any collection of geographically associated areas. The micromapST package uses the standard graphics and RColorBrewer packages to rapidly create highly readable linked micromap plots. This gives the user the ability to explore different views of their data quickly. The micromapST uses the border and name information contained in border group datasets to define the geographical areas used in creating the linked micromaps.

The main aim of redesigning the graph is to make it understandable by anyone who is not aware of the content. The linked micromap is the best package to represent the data without confusing. The below figure represents the Analysis of Incidence Rates on Stomach Cancer in USA in Between 2012-2016, For All Race, Ages and Gender. The graph gives the names of the states with their colors, confidence intervals for age adjusted incidence rates, confidence intervals on recent 5-year trend.



**FIGURE 3: LINKED MICROMAP**

ADDITIONAL VISUALIZATIONS:

By using the CDC dataset, we could implement some additional visualizations. In order to do that, we need to make some changes to the csv file and cleaning of data is required as data from USA does not include data from Puerto Rico and different functions are used to give some values which are used for additional visualizations.

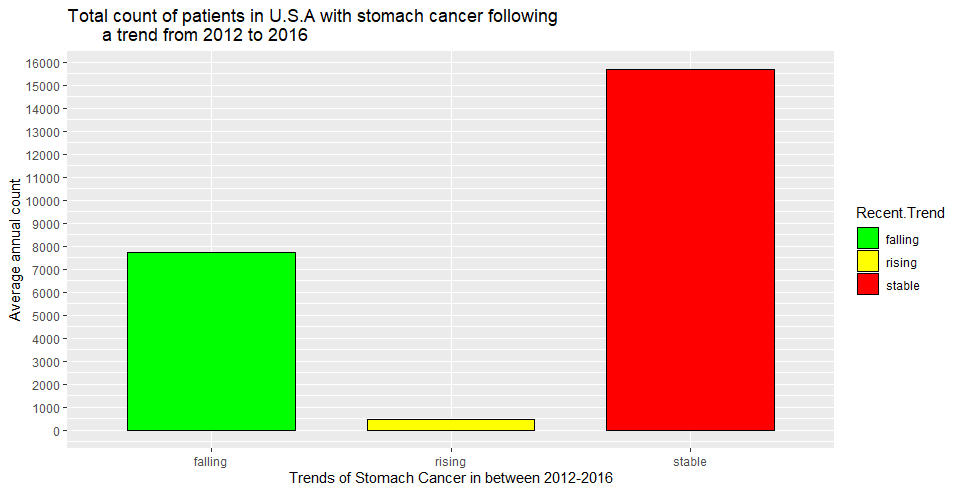


**FIGURE 4: GRAPHS REPRESENTING RATES**

The above graphs only give the information and analysis of rates of cancer by sex, age group. To get an in-depth idea of dataset, we must do additional visualizations which gives average annual count, age adjusted incidence rates in between 2012 to 2016 with a recent trend. By producing the additional graphs which helps to redesign the data with new set of plots.

The tidyverse package is installed for usage of ggplot for data visualization. The below bar chart, box plot and plot order- Row labelled plot.

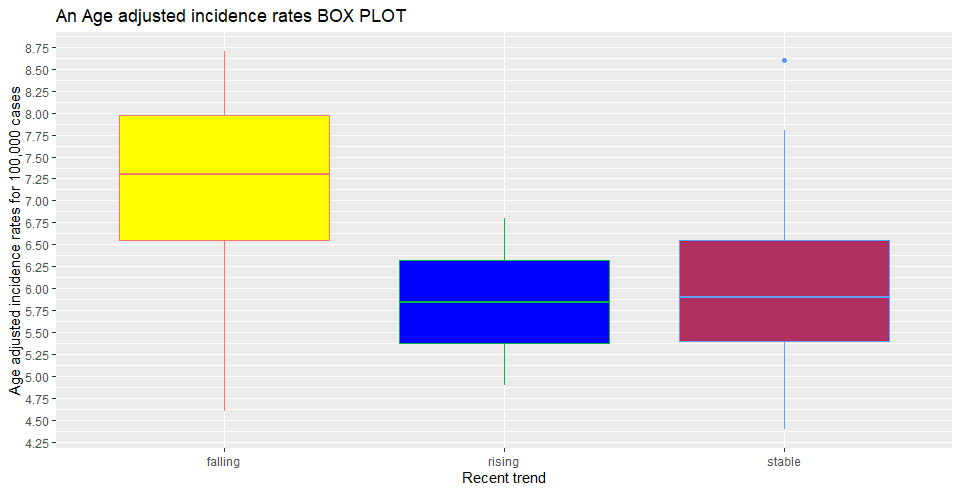
BAR GRAPH:



**FIGURE 4: BAR GRAPH**

The above bar graph represents the recent trend of average annual count of patients in USA. The overall graph represents falling, rising and stable trends from 2012 to 2016.

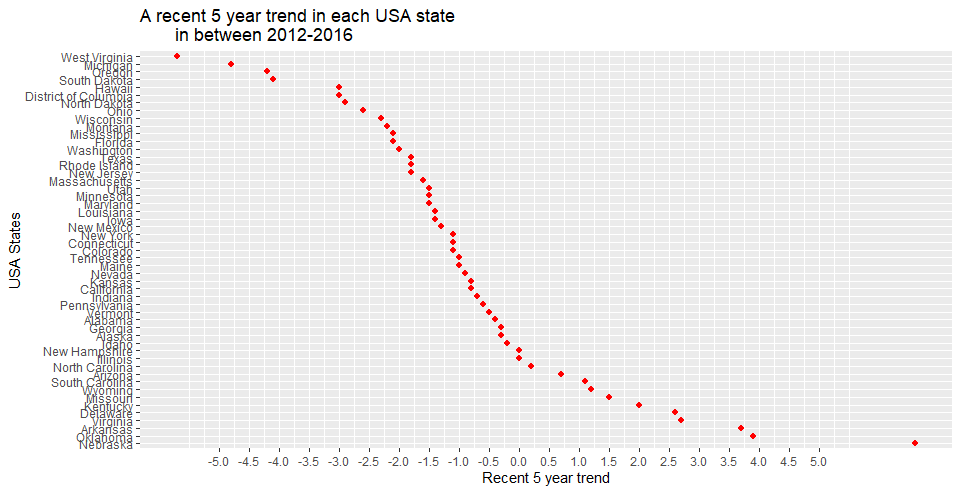
BOX PLOT:



**FIGURE 5: BOX PLOT**

The box plot gives an idea of age adjusted incident rates for 100,000 cases and here we can observe the falling, rising and stable analysis of recent trend.

ROW- LABELED PLOT:



**FIGURE 6: ROW- LABELED PLOT**

The above plot represents a five-year trend in each state in USA. These additional visualizations are implemented to get an enhanced idea on dataset by observing the graphs and plots.

**CONCLUSION:**

By redesigning the graph and adding additional visualization to analyze and get an in-depth idea. With this we would conclude with the following understandings. Initially, the geographical areas like New York, Hawaii, District of Columbia high age-adjusted incident rates whereas geographical areas like West- Virginia, Michigan, have experienced a low 5-year recent trend. Finally, we can conclude that most of the total annual count of patients in USA followed a stable recent trend.

**REFERENCE:**