STAT 515 Mid Term Project

Redesign for Bad Graphs

Group 8

Rohith Panjala

Adithi Kallem

Table of Contents

Introduction:

Statistical graphs are a powerful tool to express a large amount of information in a precise manner. But in some cases, it might deceive the audience.

Bad graphs are those that manipulate the provided statistical data to show false conclusions. Thus, bad graphs can be misleading as the users only look at the graphs not the data.

Using the data and presenting it as graphs will ensure that the data is treated fairly.

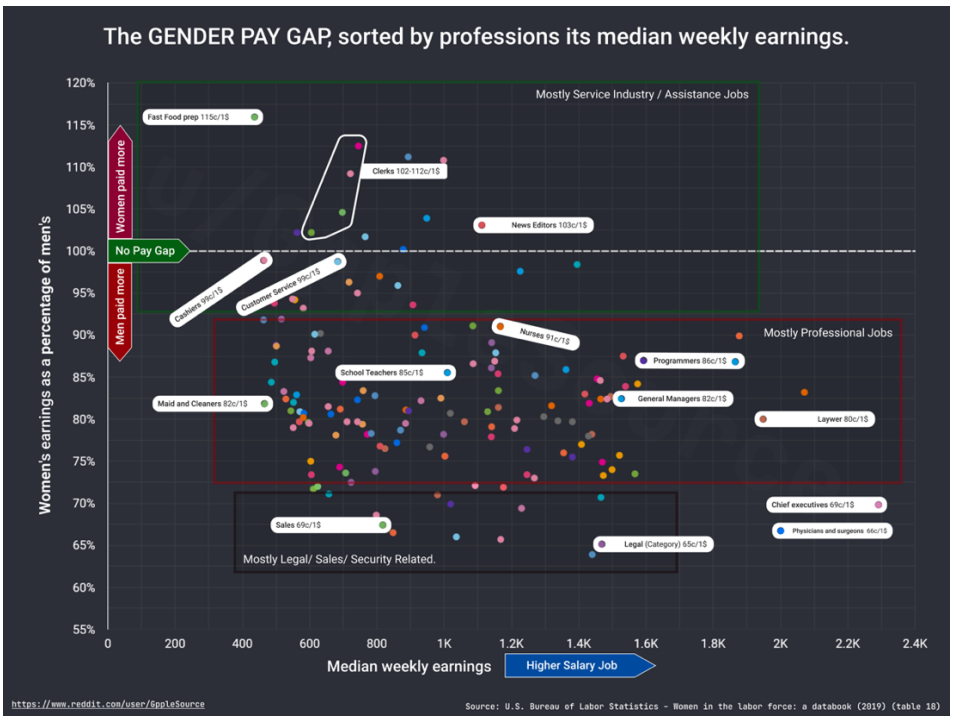
By comparing two variable quantities, graphs are utilized to appropriately interpret the data information in accordance with the predicted statics.

They are employed in the analysis of massive datasets in order to evaluate the relevant difference between the attributes considered and arrive at a predictable result or learn something. Misleading or bad graphs are a serious error that can result in incorrect predictions and interpretations when this analysis is used further.

A picture is worth a thousand words, and a graph is worth a thousand numbers only if it is clear, concise, and correct. Always graphing your data properly will often obviate the need for further analysis. Do not lose sight of the focus of the graph when you are building it.

Bad graph1:

The Gender wage gap, sorted by profession



This graph is used to represent the predictive analysis of how much more money men make than women across different profession considering weekly earnings

Source:

Graph:
https://www.reddit.com/r/dataisbeautiful/comments/p0z3z2/oc_the_us_gender_pay_gap_visualised_by/

Dataset:
https://iwpr.org/wp-content/uploads/2020/07/2020-Occupational-wage-gap-FINAL.pdf

This Visualization which concise a bad graph is taken from the reddit.com and the data for visualization is taken from iwpr.org. They have earnings for men and women in 2019 for full time workers in dollars and women's Earnings as a Percentage of Men's.

Context:

The description of the difference between wages received according to their gender across different professions is the theme of the considered visualization. They plotted women's earning as a percentage of men to the median of weekly earnings across professions, where the color of each point corresponds to the profession, creating a graph of too much information about gender income inequality across various profession to represent how much more money men make than women with the wage gap difference.

Strengths and Weakness of considered bad graph:

Strengths:

Apart from the gender wage gap, the graph also shows the sector wise segregation where it indicates the service industry has the less wage gap bias and the professional jobs has a moderate wage gap, and it also shows that the legal and sales industry has a drastic wage gap when compared to all other sectors.

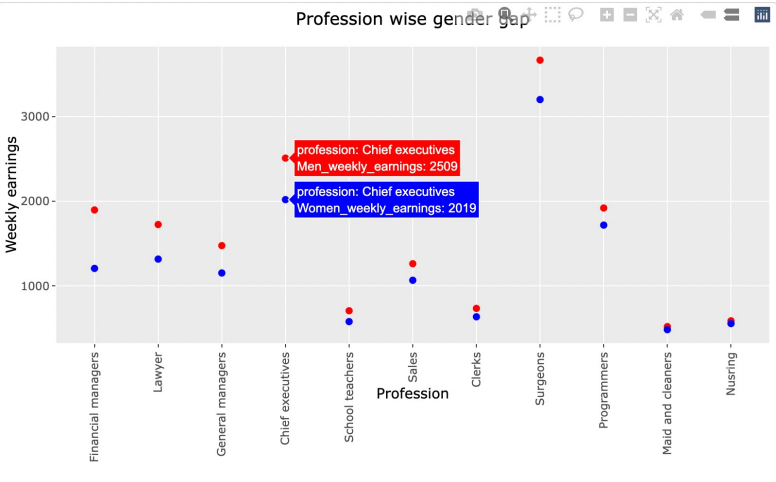
The graph also indicates the no way gap line which helps us to find the respective fields where women are getting paid more than men.

Weakness:

For the given graph, the alignment is missing as it was given the labelling wherever it wanted to be.
And the graph has too much data representation where it shows multiple scatter plots for each profession with different y coordinates.

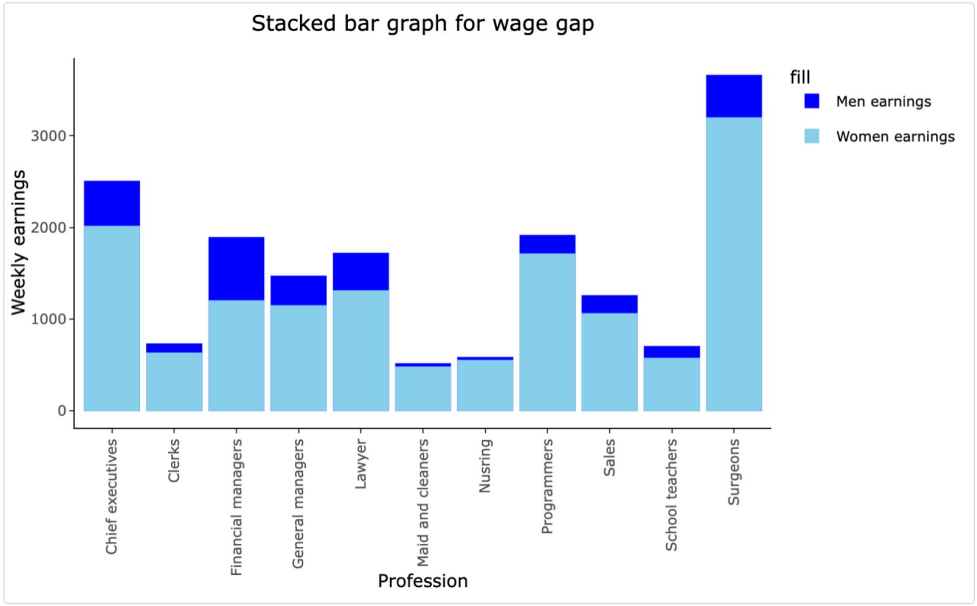
Redesigned/developed graph:

Graph 1:



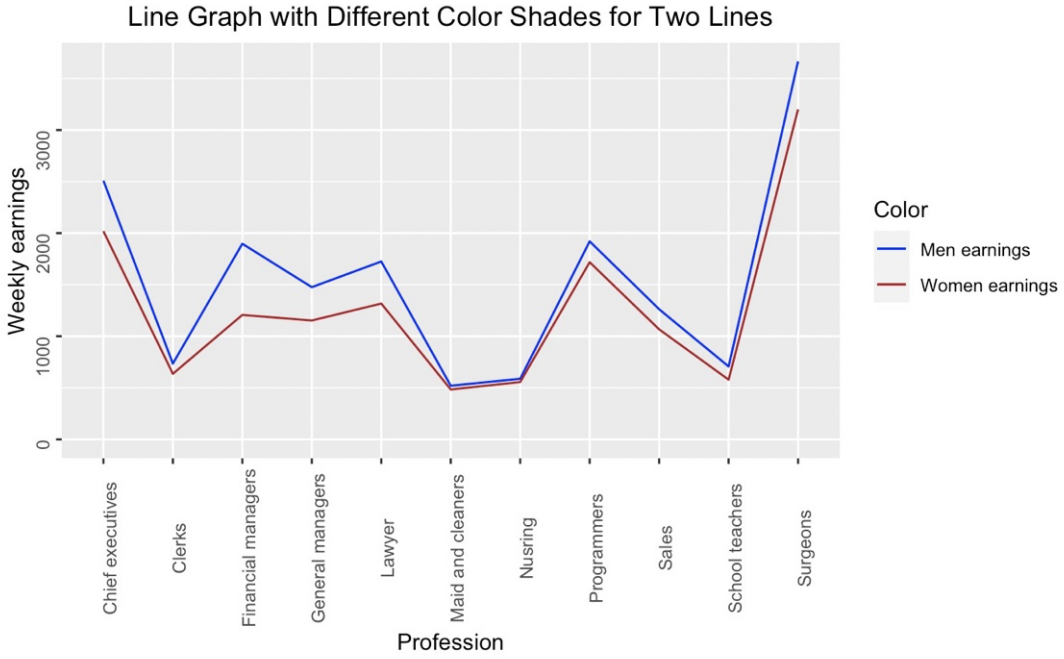
The above graph is a scatter plot which gives information about the wages of male in red and female in blue respect to the profession. The professions are sorted in ascending order with respect to women's earning as a percentage of men. We can see that profession in which the women's earning as a percentage of men is less. In the bad graph, we are not clearly able to figure out which profession had more wage differences and which profession did not.

Graph 2:



The above graph is the redeveloped graph using a stacked bar graph representation for the Input dataset of values. The bar graph represents the men and women earning for each profession. Men and women are depicted differently using appropriate coloring. A color legend is also displayed. We created an interactive graph using ggplotly() so that you could see the details when you hovered over the bars.

Graph 3:



.

The above line graph is the other way of visualizing the data. Which represents the men and women earning for each profession. Men and women are depicted differently using appropriate coloring. We can see that the instances where the lines are converging the respective profession have less wage gap and the instance where the lines are diverging the wage gap is high.

Bad Graph 2:

State wise US crime rate in 2018

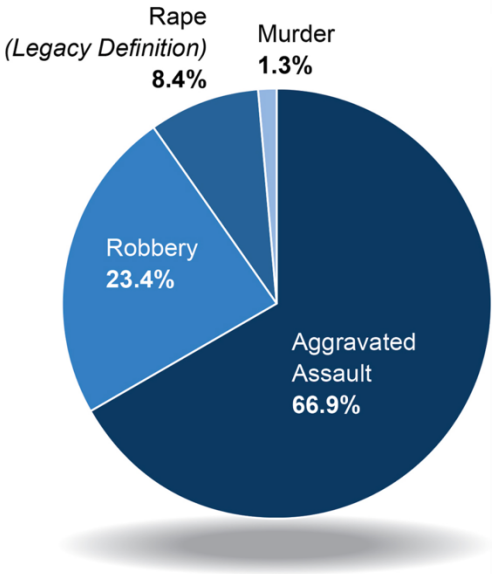
Source:

Graph:
https://www.fbi.gov/news/stories/2018-crime-statistics-released-093019

Dataset:
https://www.disastercenter.com/crime/uscrime.htm

Context:

The visualization shown above is the representation of State wise US crime rate in 2018.The graph shows the proportion of the different crimes and there rates.



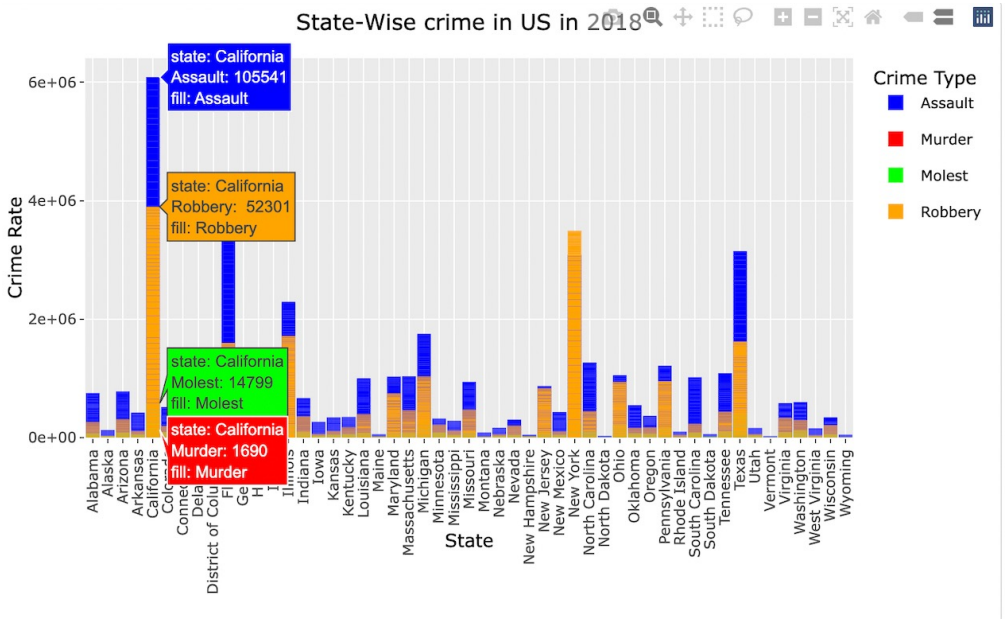
From *Crime in the United States, 2018.*

Strengths and Weakness of considered bar graph:

Strengths:
The graph is straightforward and simple to grasp. The pie chart allows the audience to quickly compare the data, of the crime rates and it makes it easy to examine the proportions.

Weakness:
In most cases, we don't use pie charts for visualizations. There is no title for the graph, but this informative data can be displayed in an interesting manner and can yield interesting insights. Even when comparing the data is simple, the results could still be inaccurate.

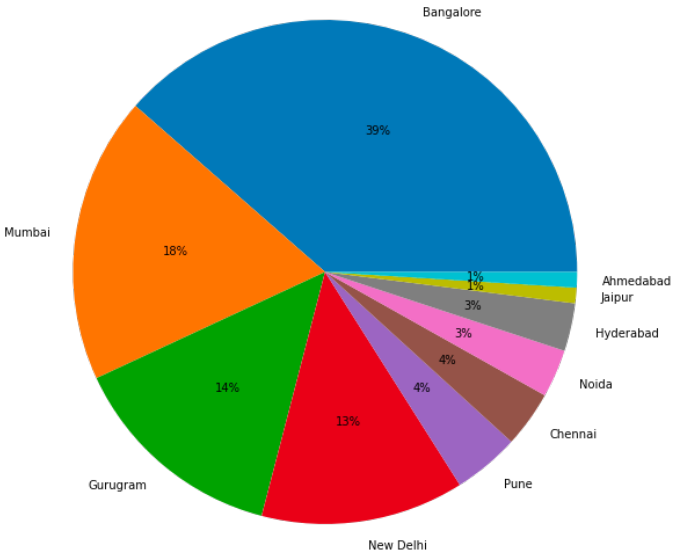
Redesigned/developed graph:



The visualization above represents the crime type across all the US states and the number of crime cases. The stacked bar graph is made interactive using ggplotly() so that you could see the finer details when you hovered over the bars. The graph clearly show that the number of cases for Assault and Robbery are comparatively high for each state. When we hover on the bar graph, we get the number of cases for each crime type.

Bad Graph 3:

Indian startup companies statistics 2021



Source:

Graph:
https://startuptalky.com/indian-startup-funding-investors-data-2021/
The data used to create the pie chart above was taken from Startup Talky and plotted using ggplot2 on the R platform.

Dataset:
https://startuptalky.com/indian-startup-funding-investors-data-2021/

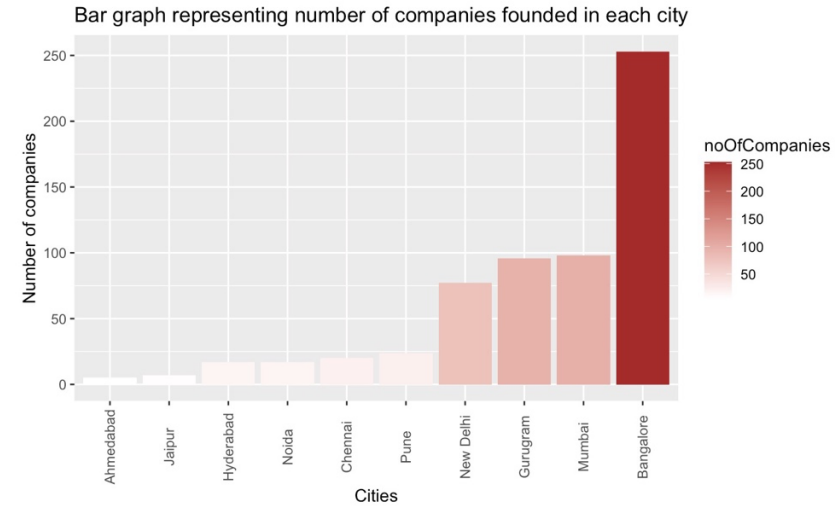
Context:
The primary goal of the visualization shown above is to illustrate the proportion of startup companies that have received investment across various Indian cities. In the graph above, the number of startups in each city is shown as a percentage.

Strengths and Weakness of considered bad graph:

Strengths:
It's simple and easy to understand the graph, The pie chart enables the audience to see the data comparison at a glance where we can see the proportion for each city clearly.

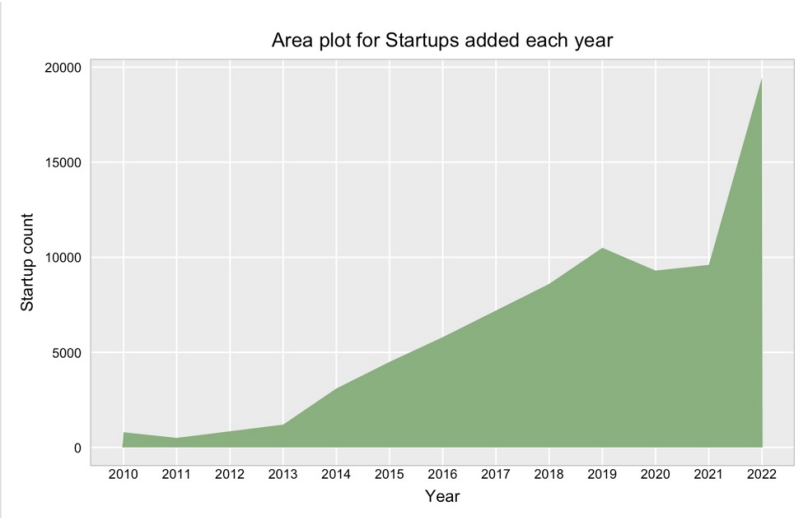
Weakness:
We don't use pie charts for most of the visualizations. This informative data can be represented in most interesting way and can extract interesting insights and there is no title for the graph
Even if the data comparison is easy, sometimes there might be incorrect conclusion.

Redesigned/developed graph:
Graph1:



In order to display the number of startups formed in Indian cities, the graph has been modified. We changed the pie chart from percentages to a bar chart so that we could simply count the number of startup businesses that were established in each city. In addition, we added a color gradient that reflects how many companies are in each city.

Graph2:



The above visualization is a representation of number of startup companies established in India for that particular year. From the year 2010 -2013 the growth of startups is very passive and constant, from 2013 there is a linear growth till 2019. As the pandemic hit nationally and globally the push for establishing the new startups has seen a down trend. Later from 2021-2022 the growth of establishing new startups has been doubled.

Conclusion:

A picture is worth a thousand words, and a graph is worth a thousand numbers only if it is clear, concise, and correct. Data Visualization can be good or bad. Visualization is beneficial only when the chosen data type fits the data being shown. The message that is to be delivered to the viewer needs to be considered when designing a visualization. The above bad graphs are confusing, and they do not deliver the necessary information accurately. We redesigned the bad graph into a new interpretation.