DATA VISUALIZATION

A PORTFOLIO OF DATA VISUALIZATIONS

JJ Zhang

FALL 2024

CITY COLLEGE SAN FRANCISCO

INTRODUCTION

Welcome to my data visualization portfolio! My collection showcases a variety of visualizations designed to inform, explore, and tell compelling stories with data. Drawing inspiration from Edward Tufte's principles, most of my designs emphasize clarity, simplicity, and the effective use of data. One of the featured visualizations is created using Python, demonstrating my technical expertise in blending programming with design to bring data to life. Through these works, I aim to present information in ways that are both insightful and aesthetically compelling. Through these projects, I aim to blend clarity, creativity, and technical precision to make data more accessible and impactful.

OVERVIEW

This portfolio demonstrates multiple different data visualizations with different goals:

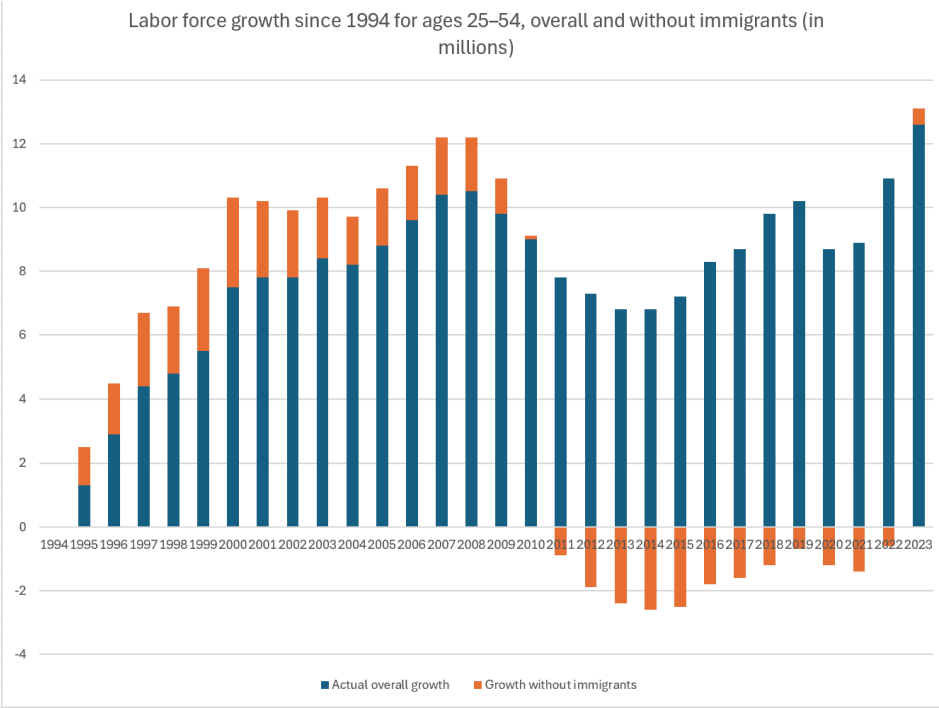
1. Graphical Integrity
2. Visualization Improvement
3. Visual Impact
4. Visualizations and Programming
5. Identifying Relationships

Graphical Integrity

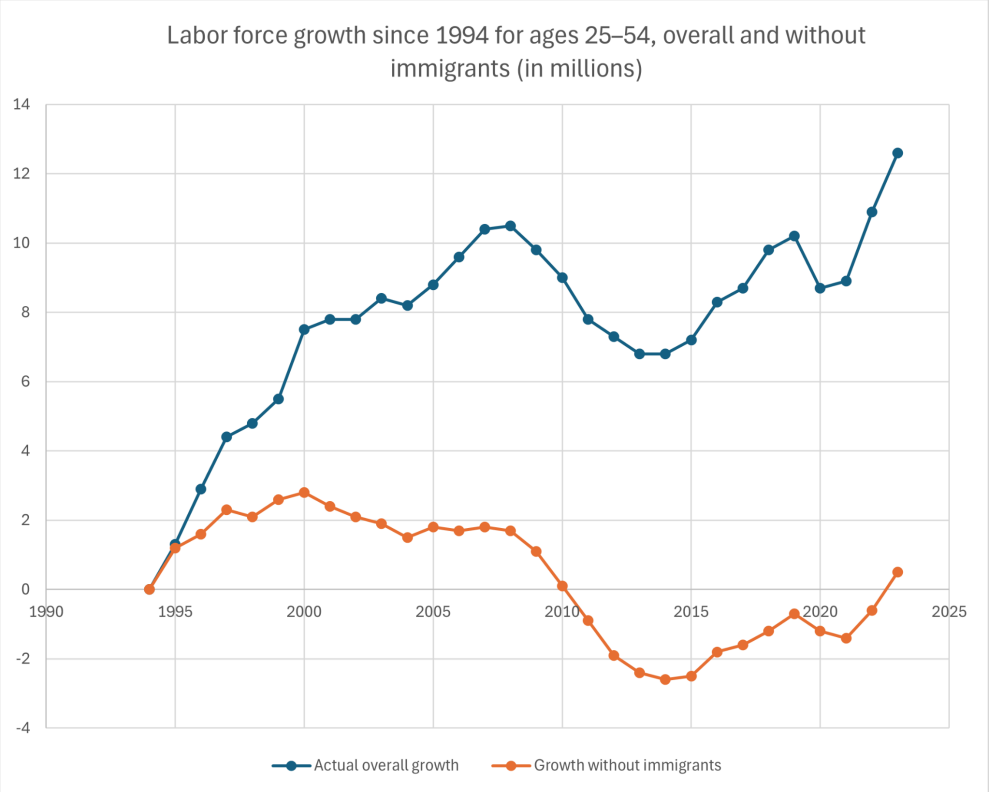
Intention Statement: Using a dataset from the Economic Policy Institute, I intend to create a graph that demonstrates the essential role immigrants play in labor force growth in the United States. The dataset contains actual labor force growth and what the labor force growth would look like if immigrants were excluded, illustrating that immigrants have been a significant factor in sustaining and expanding the U.S. workforce since 1994. My goal with the graph is to convince viewers that immigration brings a positive impact into the economy by contributing to a robust labor force. The source is: https://www.epi.org/publication/u-s-benefits-from-immigration/#epi-toc-6.

Tell a Lie

First, an example of a poor visualization that hides this truth. With the chart below, I used a stacked chart to induce the viewer to think that the growth without immigrants in the labor force would be higher than the actual overall growth, when it's not the truth. The orange portion represents the total growth without immigrants for the particular year, not how much higher/lower than actual overall growth as the audience would be induced to think when viewing this chart. This graphic tries to convince the viewers that immigration brings a negative impact into U.S. labor force growth.



Here we see a much improved image, with a line graph that allows for a comparison of the actual labor force growth with vs. without immigrants over the years. Visual clutter is reduced with evenly spaced axes. Each line uses a different color that highlights the trend over the years. I used Google sheets to edit the data and create the graph. I also used Microsoft Paint to resize the chart for best size ratios.



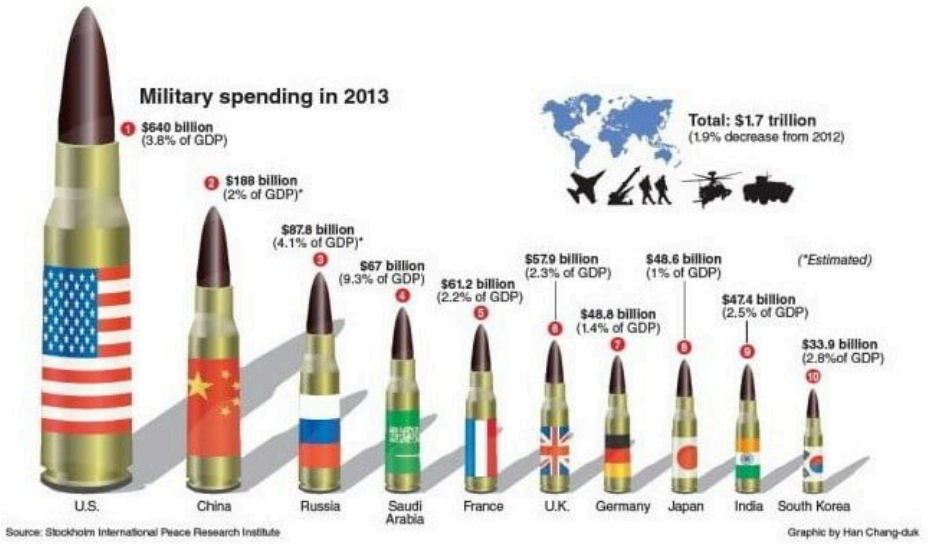
Visualization Improvement

The graphic from the Korea Herald news article Global military spending drops in 2013 led by U.S. cut (https://www.koreaherald.com/view.php?ud=20140415000892#) displays a chart showing military spending in 2013 by various nations, using bullets to represent each country's spending levels. The graphic contains the monetary value of each country's military spending as well as their respective GDP percentages. Although the graphic is visually appealing with bullets representing each nation, the graphic fails in several areas when evaluated by Edward Tufte's standards. To make the chart more effective, I recreated the chart based on Tufte's principles and other best practices for data visualization as below.

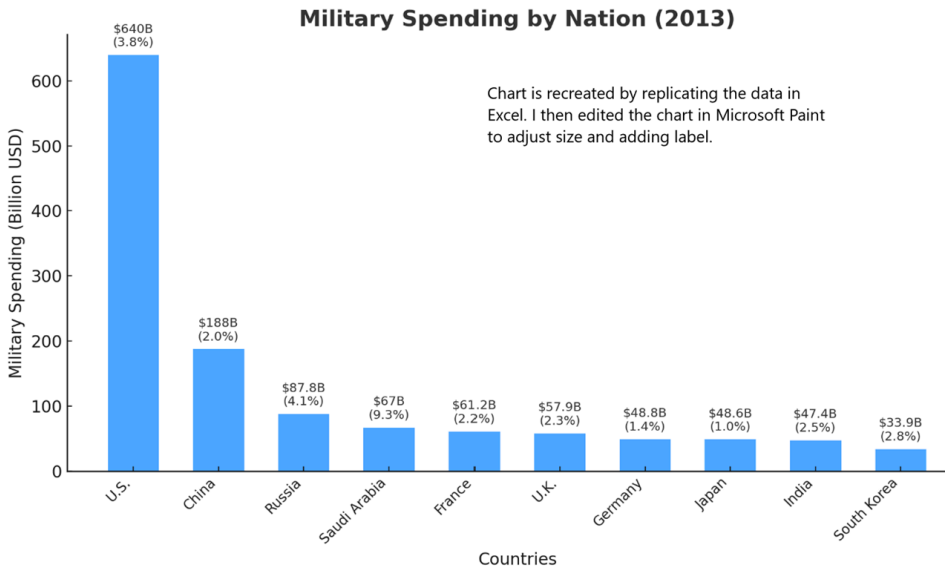
- Graphical Integrity- In recreating the chart, I replaced the bullets with bars in a bar chart, which allows for precise comparisons of military spending across countries. In the original chart, the size of each bullet is not a consistent representation of the relative spending. For example, the size difference between China's $188 billion bullet and the U.S. 's $640 billion bullet is not to scale, potentially misleading the viewer about the relative spending levels. By using a more direct, simplified bar graph, I am able to display the dataset more accurately without the unnecessary embellishment.
- Data-ink ratio- In the chart I recreated, I avoided non-data ink (e.g., bullets or decorative icons on the original chart). I also attempted to use the chart space efficiently by evenly spacing out the bars on the bar chart. This aligns with Tufte's advice to minimize redundant data-ink and maximize the impact of actual data.
- Presentation of information- The original graphic lacks clear axes and data labels, which makes it hard for viewers to compare categorical data. As a solution, I added labelled axes as well as GDP percentage labels above each bar. I believe this placement allows viewers to interpret the monetary and proportional (GDP) information in a single glance.

On the next page is a snapshot of the graphic in the original article and my recreated, improved graphic.

Original Graphic



Improved Graphic



Visual Impact

I have created a word cloud below with several deliberate design choices:

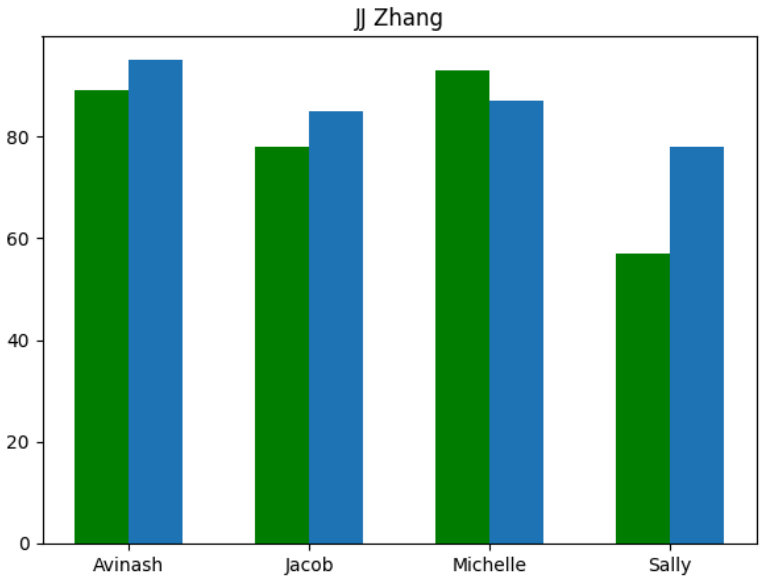
- **Shape Design**: I used a silhouette of Sherlock Holmes, which adds narrative depth and visual interest. The shape ties directly to the theme, making the visualization both informative and engaging.
- **Color Palette**: I used contrasting colors—light pink and white on a dark blue background to create a strong visual appeal while maintaining legibility. The colors are thematic, possibly evoking a classic and mysterious mood associated with Sherlock Holmes.
- **Font Size Scaling**: Larger, more prominent words like "Holmes," "upon," and "one" reflect higher frequency, directing attention to the most repeated terms.

Data source: [The Adventures of Sherlock Holmes]

Tools: [wordclouds.com]

Visualizations and programming

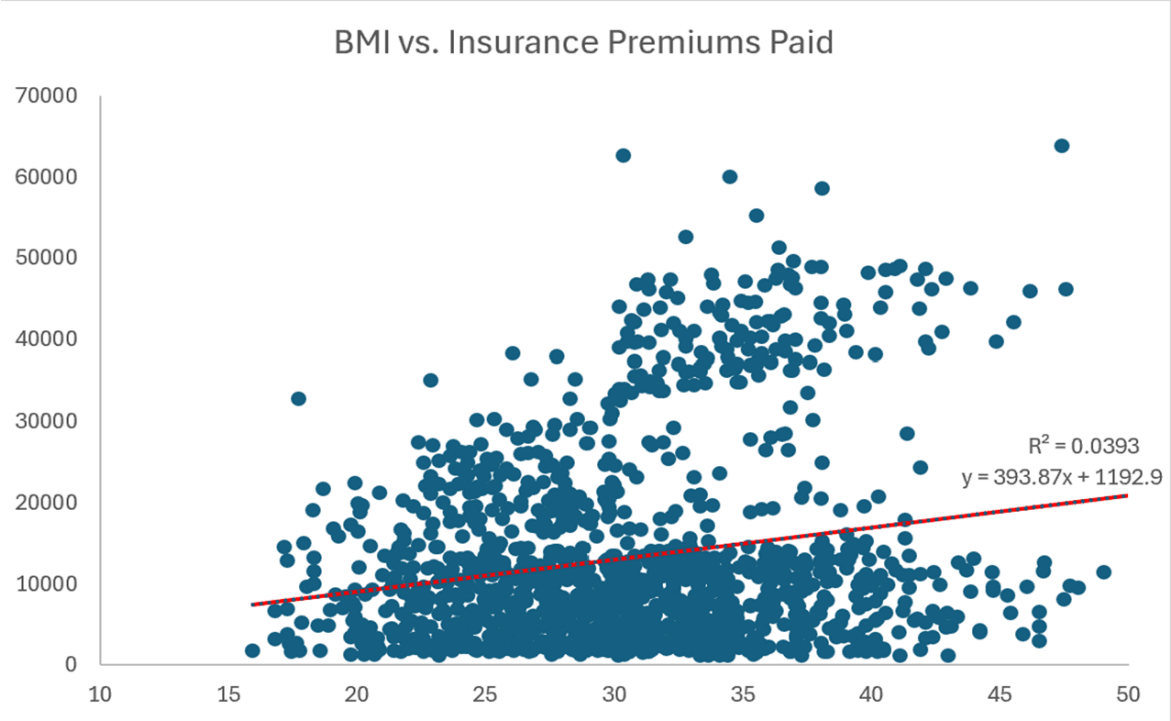
```
import numpy as np
import matplotlib.pyplot as plt
figure = plt.figure(figsize = (7,5))
names = ['Avinash', 'Jacob', 'Michelle', 'Sally']
scores = [89, 78, 93, 57]
scores2= [95, 85, 87, 78]
positions = [0,1,2,3]
positions2= [0.3, 1.3, 2.3, 3.3]
positions3= [0.15, 1.15, 2.15, 3.15]

plt.bar(positions, scores, width=0.3, color='g')
plt.bar(positions2, scores2, width=0.3)
plt.xticks(positions3, names)
plt.title('JJ Zhang')
plt.savefig("grades.png")
plt.show()
```



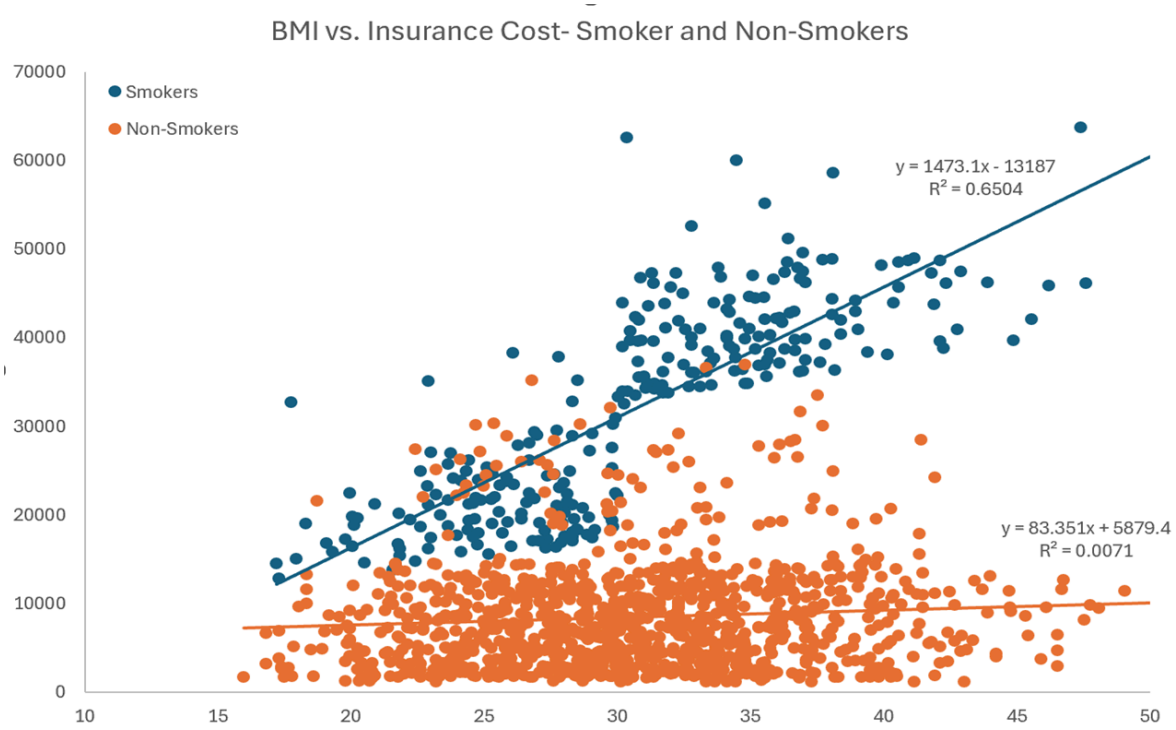
Identifying Relationships

With the below two scatter plots, I plan to explore whether there's a correlation between people's BMI measure and the cost they pay for health insurance. The data set I found can also further explore the difference between smokers vs. nonsmokers' BMI levels and their respective insurance costs.

I imported the data from a Github dataset. The link to the data source is: https://github.com/eajitesh/insurance-linear-regression-example. I used Microsoft Excel to analyze the data and create the scatterplot. The scatterplot is edited via Microsoft Paint to reach the golden ratio of size.



The scatterplot above suggests that body mass index (BMI) and insurance charges are positively correlated. Customers with higher BMI typically tend to pay more in insurance costs.



This scatter plot shows that while nonsmokers tend to pay slightly more with increasing BMI, the smokers pay higher insurance costs overall.

To further emphasize this fact, I have added two regression lines, corresponding to smokers and nonsmokers. The regression line for smokers has a much steeper slope, indicating the positive relationship between bmi and increased insurance costs.