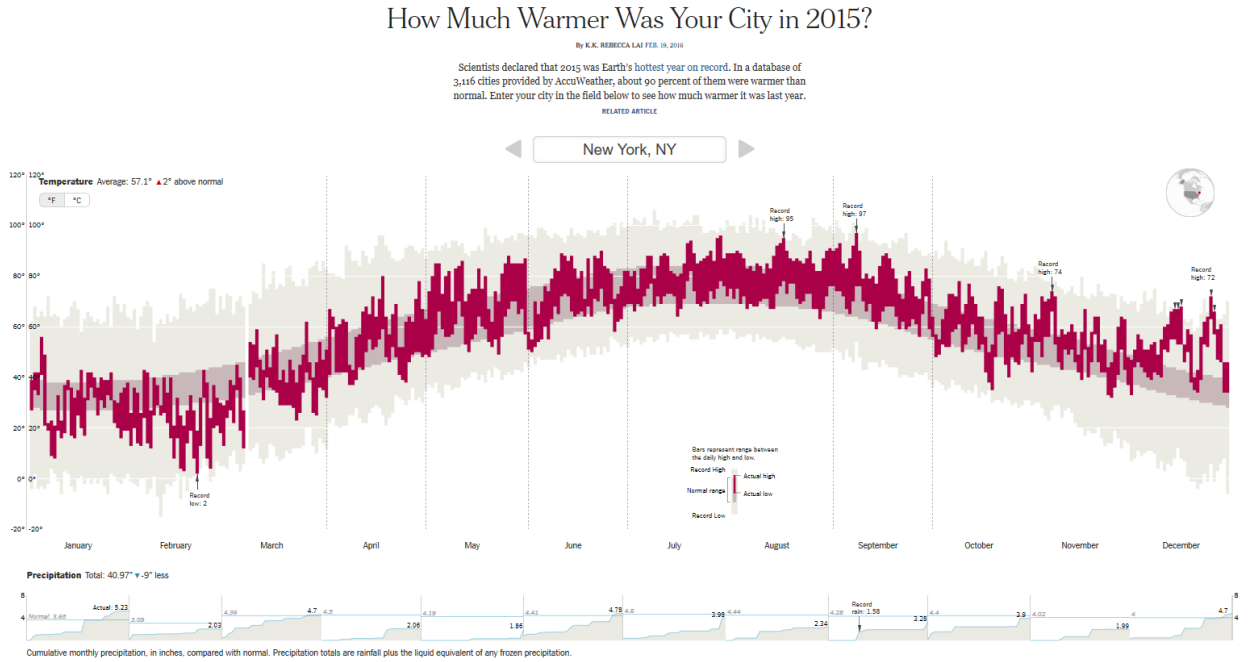
Assignment 2: Visualization in Media and Literature

1. Interactive Visualization

Lai, K. k R. (2016, February 19). How Much Warmer Was Your City in 2015? *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2016/02/19/us/2015-year-in-weather-temperature-precipitation.html#new-york_ny

On February 19, 2016, The New York Times published an interactive chart showing detailed temperature and precipitation patterns for 1,801 American cities and 1,334 other locations around the globe for 2015. A screen shot of the chart for New York City is shown below.



The chart basically consists of two plots juxtaposed, one above the other, sharing the same x-axis of 12 months. The top plot provides the actual daily high and low temperatures in 2015, the normal high and low temperatures from 1981 to 2010, and record high and low temperatures. The bottom plot consists of 12 mini-plots, one for each month, providing the cumulative precipitation for the month and normal precipitation lines for each month.

As an interactive chart, users can type in the city name in the top middle box to take a look at its temperature and precipitation. There are thousands of cities whose data are included in the database. Besides, users can choose the temperature unit as either Fahrenheit or Celsius. In addition, there is a small picture of the Earth on the top right of the chart. The red dot generally represents the location of the city that the user selects.

There are some excellent details that could be found in the two juxtaposed plots.

In the top plot:

- Three colors are used for the record temperature, normal temperature, and daily temperature. They look like three laces visually. Users can clearly see the distribution and trend of the temperatures in a year.

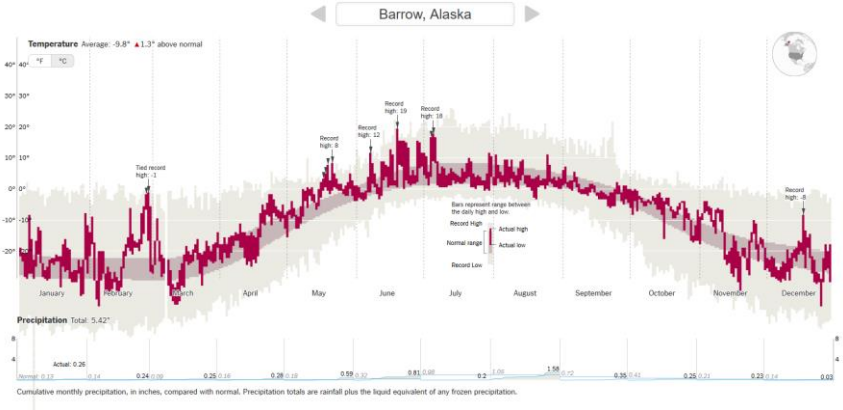
- A vertical dotted reference line appears in order to separating each month.
- Reference text is included for those days that tied or exceeded the record high/low.
- The y-axis tick marks are at 10-degree intervals.
- There is no x-axis, axis label, or tick marks, tick mark labels along the bottom of the plot. Instead, the months are included along the top axis.
- There is a legend that indicates how to read the actual, normal, and record temperatures.
- Each high and low for each day (whether actual, normal, record) is represented as a narrow rectangle.
- The average temperature of the year 2015 and the variation compared to normal are demonstrated on the left top.

In the bottom plot:

- There are 12 small plots, and each provides a cumulative precipitation for each month.
- Each has a reference line for the normal precipitation.
- The area below the cumulative curve is shaded.
- The x-axis is the mirror of the top plot.
- The total precipitation of the year 2015 and the variation compared to normal are demonstrated on the left top.

Suggestions for improvement:

- For some cities, the chart does not show the record high and low temperatures, such as Seattle, Washington. I hope more data can be collected.
- For some cities, the temperatures or precipitations may be too low or too high that the general range in the chart is not appropriate for that. So I suggest adopting adaptable and alterable ranges of temperatures and precipitations for different cities. For instance, the screen shot below shows that there are no accurate tick marks for low temperatures.



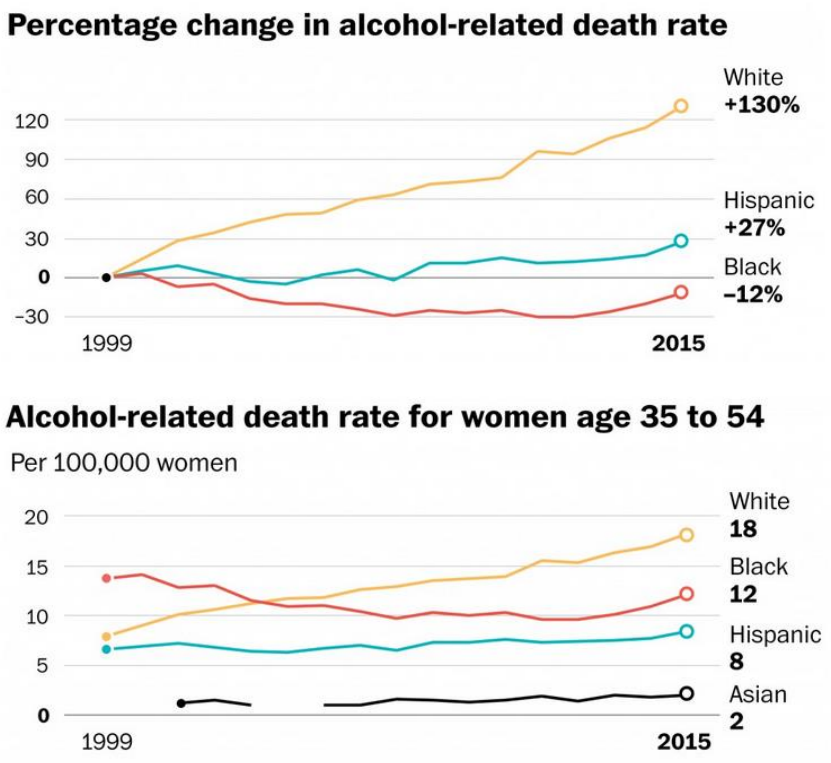
- When clicking the left or right arrow button close to the box, it switches to another city that is before or behind the current city in alphabetic sequence. I find it not reasonable. I suggest grouping cities by geographical regions, and then sorting in alphabetic sequence in each small region.

2. Visualization for Comparative Study

Keating, D. (2016, December 23). Nine charts that show how white women are drinking themselves to death. *The Washington Post.* Retrieved from https://www.washingtonpost.com/news/national/wp/2016/12/23/nine-charts-that-show-how-white-women-are-drinking-themselves-to-death/

The study aims to find why middle-aged (age 35 to 54) white women are dying more often even while death rates for other groups continue to go down. This report demonstrates that a simple answer is that they consume a lot more drinking.

There are nine charts showing different evidence for this finding. I pick the first two relevant charts as an example to analyze its performance of visualization.



The first chart shows the trend of percentages changes from 1999 to 2015 in alcohol-related death rate for three groups of people (the middle-aged white, Hispanic and black women dying of too much alcohol).

The second chart shows the trend of numbers of women dying of alcohol per 100,000 women for four groups (adding an Asian group) from 1999 to 2015.

For the two plots:

- The purpose of both charts is to demonstrate the trend of different groups and the general comparison between them. Therefore, concise line graphs are good in terms of intuitive visualization.
- Both charts do not emphasize accurate values during this variation, instead, they just show the accurate values of the final year. It is reasonable since readers do not need to know the accurate values of each year. The trend line itself matters most. It can also make the comparison clear.

- Both charts utilize colors of high contrast to show different groups, making it easy for readers to compare between them. The color for the same group also remains the same in different charts.
- The two charts should be observed together to make a better comparison.

For the first chart:

There are some descriptions in its original text. “In 1999, white and Hispanic women had relatively similar rates of death from alcohol, and the rate for black women was considerably higher. But since then, the death rate for blacks has gone down, the death rate for Hispanics has gone up a bit, and the death rate for white women climbed 130 percent.” The accurate alcohol-related death rates for middle-aged women of different groups have been shown in the second chart.

The chart assumes the starting rate for all groups is zero. Then the final increasing percentages for different groups are calculated based on the initial values for different groups.

For the second chart:

The data about Asian group is added in the chart using black line.

Suggestions for improvement:

- There is a break point in the black line for Asian women in the second chart. I have no idea what it implies. I suggest providing some explanation for that.
- For the first chart, the zero starting point for all groups may not be a good choice. I consider that the base values in 1999 for different groups are different, so the direct comparison of increasing rates of different groups may generate some bias.
- For the second chart, the data exist randomness and inaccuracy since they are acquired for selected 100,000 women. Compared with the first chart, we may find non-correspondence in detailed values. For instance, the increasing rate for Hispanic is (8-6)/6 = 33.3% based on the second graph, while it is 27% in the first graph.

3. Statistical Visualization

Center, R. (2016, October 27). High School Benchmarks 2016: National College Progression Rates | National Student Clearinghouse Research Center. Retrieved from https://nscresearchcenter.org/high-school-benchmarks-2016-national-college-progression-rates/

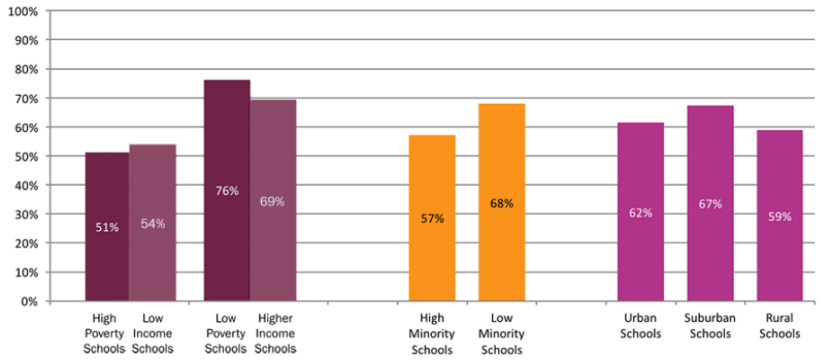
This is not a nationally representative sample of schools or of high school graduates. Compared to all U.S. high schools, participants tend to have greater representation among schools with more low income students, more minority enrollments, and more urban locales.

The dataset used to create the graphs in this report is accessible, so I choose one of the graphs to recreate. The dataset is saved as Appendix-C.xlsx file, containing multiple tables for generating the graphs.

There is an excellent stacked horizontal bar chart that represents the distribution of fields of study among STEM completers of class of 2009. I tried to recreate it utilizing matplotlib library in Python but failed finally. Instead, I choose another bar chart to recreate.

As for recreating visualization, I choose Figure A in this report as the target graph, as is shown below.

Figure A. College Enrollment Rates in the First Fall after High School Graduation, Class of 2015, Public Non-Charter Schools



This above figure shows the rates of immediate college enrollment in the first fall after high school graduation for the class of 2015. Income was the strongest correlate with immediate college enrollment. Students from higher income schools were more likely to enroll immediately than students from lower income schools (69 percent and 54 percent, respectively). The gap became even larger when we examined this outcome for graduates of high-poverty schools and low-poverty schools. A 25 percentage point difference exists between high- and low-poverty schools (51 percent and 76 percent, respectively).

The percentage of minority students at schools was also a strong correlate. Students from low minority high schools were more likely to enroll immediately than those from high schools with higher minority populations (68 percent and 57 percent, respectively).

While, location was not as strongly correlated, but still demonstrated some relationship with immediate college enrollment. Students from suburban schools (67 percent) were more likely to immediately enroll than those from urban (62 percent) or rural (59 percent) schools.

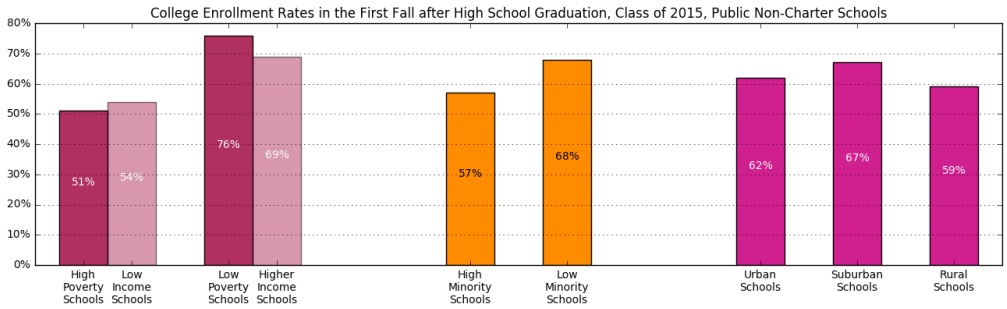
The figure generally clusters the bars in three groups, aiming at making comparisons in three aspects. Income, minority and location issues are three factors that may be correlate with the college enrollment rates. Different colors are also assigned for different groups of bars. The meaningful grouping makes the figure a good graph for analysis.

Suggestions for improvement:

- The highest value in the y-axis is 100%, which causes some white space on the top of the figure. I suggest reducing the range of y-axis.
- Maybe adding some annotations (text and directed lines) can help better understanding how to make comparisons and what analytical results can be found based on the bar chart.

My re-creation graph:

I used some data in Table 1 of the file Appendix-C.xlsx.



- I set the x positions of bars manually, and I am not sure if there is a better way for positioning.
- I try to retain its original grouping since it can help readers to make comparisons easily.
- I add "\n" manually to each string of the x-labels to avoid overlapping.
- I retain the text of percentage in the middle of each bar, and set its color to black or white depending on its back color of the bar. If the percentages are displayed on the top of the bars, readers will have to move their eyes above instead of focusing on the bars themselves.

Differences compared with the original graph:

- I did not modify the max value of y-axis to 100%. Instead, the graph generated its adaptable y-limit automatically. In this case, the max value is 80% in y-axis.
- I was wondering why the edge colors are still black even though I already set them to "none".

4. Geographical Visualization

Wang, L., & Chen, L. (2016). Spatiotemporal dataset on Chinese population distribution and its driving factors from 1949 to 2013. *Scientific Data*, *3*, 160047. https://doi.org/10.1038/sdata.2016.47

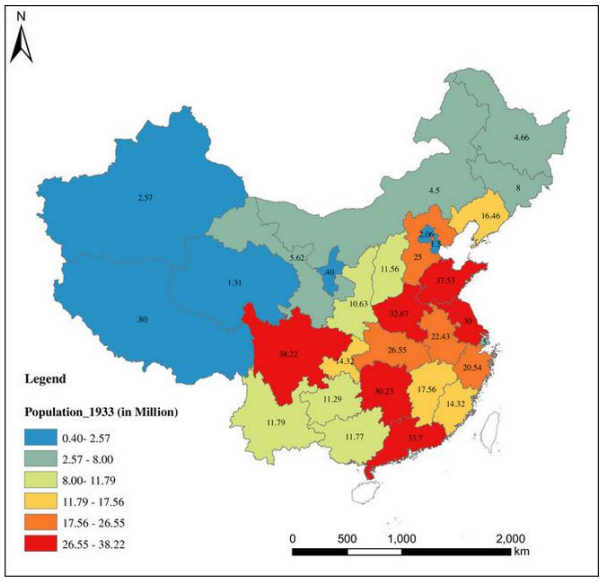
The academic article presents a dataset on Chinese population distribution and its driving factors over a remarkably long period, from 1949 to 2013.

"With respect to the spatial distribution of China's population, the pattern reflects a great difference between the southeast and northwest parts of the country (Figs 5,6,7,8). Most of the population is concentrated in the southeastern side of China. Guangdong has been the most populous province with a population of 106.44 million in 2013, while Tibet is the least populous province with a population of 3.12 in 2013. The spatial distribution of the population has not changed much over the eight decades from 1933 to 2013."

From Figs 5,6,7,8 below, we can find that the rates of population growth varied in different provinces. Showing the exact amount of populations for different provinces on the geographical map is an intuitive way to observe the distribution. Besides, the four graphs together can help make comparisons between provinces not only in the same year but also across years to find the tendency.

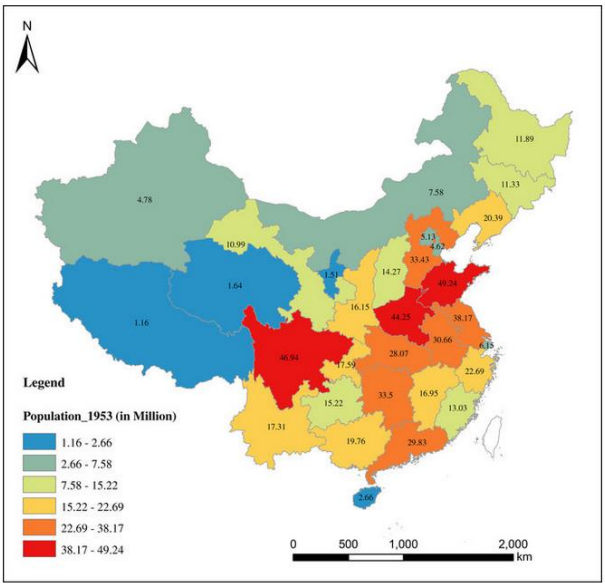
The measuring scale on the map and the North arrow also make the plot accurate and reliable.

Figure 5: Population Distribution of China in 1933 (Taiwan and Hainan are excluded).



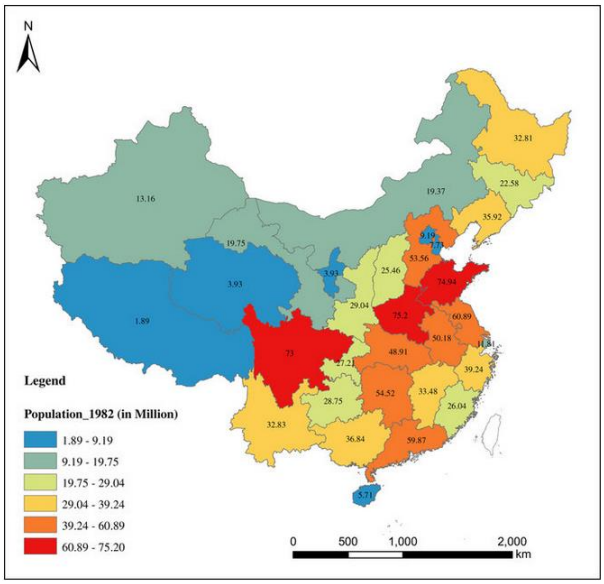
Different value of population is expressed by different color group and population of each province is labeled on the map.

Figure 6: Population Distribution of China in 1953 (Taiwan is excluded).



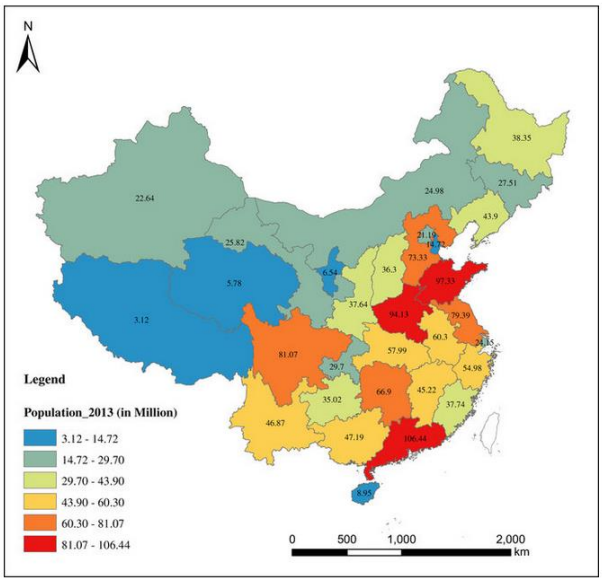
Different value of population is expressed by different color group and population of each province is labeled on the map.

Figure 7: Population Distribution of China in 1982 (Taiwan is excluded).



Different value of population is expressed by different color group and population of each province is labeled on the map.

Figure 8: Population Distribution of China in 2013 (Taiwan is excluded).



Different value of population is expressed by different color group and population of each province is labeled on the map.

Suggestions for improvement:

- For the legend, the partitions of the amount of population are different in the four years. The population is generally increasing so the detailed amount of sections for the legend should also change. However, how to set the thresholds for different colors in the legend is not clear. The comparison among different provinces in the same year is good, while the comparison between

different years for the same province only based on its variation of color in the graphs may not be quite reasonable. We can only get a general understanding of the variation of the distribution of China from 1933 to 2013.

- I suggest showing the total amount of population of China for each year. In this way, readers can coarsely estimate the percentage of any province contributing to the total population.

5. Scientific Visualization

Ikram, M. T., Butt, N. A., & Afzal, M. T. (2016). Open source software adoption evaluation through feature level sentiment analysis using Twitter data. *Turkish Journal of Electrical Engineering & Computer Sciences*, *24*(5), 4481–4496. https://doi.org/10.3906/elk-1502-173

According to previous analysis of the results, the article finds that functionality, cost saving, reliability, and community support are the most commonly discussed features.

Figure 3 shows the opinion frequency graph of different adoption factors. The x-labels are the names of the features, and y-axis reflects the amounts of papers. Each bar represents the distribution of negative and positive papers for papers with a certain feature. For instance, there are 250 positive papers and 125 negative papers in the 375 papers with functionality feature.

It can also be found that the probability or likelihood to submit positive opinions is greater as compared to negative comment for all papers with certain features.

I suppose it is a stacked bar plot. The colors of high contrast are excellent in the graph, making it easy to find distribution of positive and negative papers.

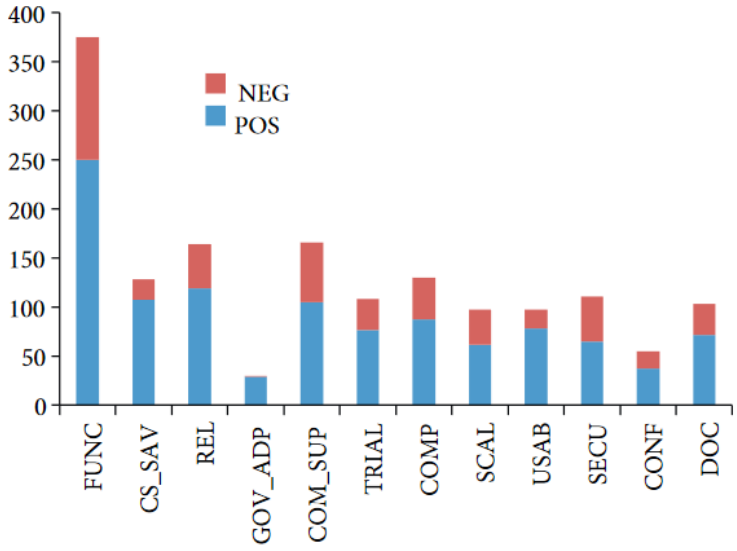


Figure 3. Opinion frequency graph of adoption factors.

Suggestions for improvement:

- Maybe the percentages of negative and positive papers for each bar could be added on the bars. But the graph should be larger and the width of each bar should also be larger in this way.
- Maybe the bars could be ranked in a decreasing order of the percentage of negative papers.