Midterm Redesign Project

Group 1: Long Zhang, Mohammad Ridwan, Baizhong Hou

1 Introduction:

As so far, we have learned graph designing via {ggplot2} and {micromapST} packages, and data preparation via {tidyr} and {dplyr} packages in R in course STAT515. This report will redesign two different types of bad graphs via all of above packages. Specifically, one graph is a complicated line plot redesign and another one is the bar plot with geospatial patterns redesign. This report provides comprehensive description and analysis for the backgrounds of original graphs, the main goals and process of redesigns, the result and conclusion, challenges, and lessons learned.

2 Redesign 1

2.1 Introduction:

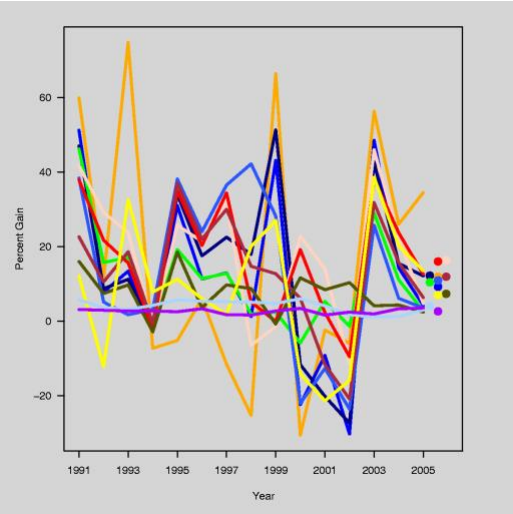


Fig. 1 Bad Graph 1

This bad graph is collected from one post on the forum of statmodeling at Columbia University (Price, 2006). It consists of 12 trends of the annual percentage gains for 12 funds in the Vanguard Group from 1991 to 2005. However, this graph performs so jumbled that we cannot figure out any valuable information. Specifically, this report will analyze the cons of this graph from the following 4 aspects:

- **Enable Accurate Comparisons:** First of all, we can see that all 12 lines are overlapping in this graph. Therefore, these complicated contents can easily distract us, which makes us hard to make further analysis and comparisons. Even there is no grid lines to help us have a good comparison in large scales.

- **Simplify Appearance:** As mentioned above, there exist so many perceptual groups at such small scales, which provide audience with a complicated sense.
- **Interpretation:** This graph lacks several important elements for conveying information, like the legends for different lines and title for the description. Without legends, this graph cannot convey any meaning and valuable information to audience, just some colorful lines. Without unit, we cannot have an accurate measure for the value. Without title, audience will not know what the purpose of this graph is.
- **Engage the Reader:** All limits above make it hard to attract readers to analyze what we can obtain from this graph. Besides these, the color allocation includes some similar colors, like blue and dark blue, orange and pink, which produces a bad visual effect.

2.2 Redesign Goals:

The purpose of this bad redesign case is to improve above cons via separating it to different perceptual groups with a scientific standard, adding necessary context, and adjusting attractive displays. The final result should satisfy and balance all of above four standards, accurate comparisons, simplify appearance, good interpretation, and engaging readers.

2.3 Redesign Process:

The main tools used in this work are R studio, {tidyr} and {dplyr} packages for data preparation, {ggplot2} and {RColorBrewer} and {gridExtra} packages for plotting graph. Specific steps are shown as follows:

2.3.1 Data Preparation

To satisfy the dataset schema used in ggplot2 package, we should transform the data schema first. To this end, call `gather()` function in package {tidyr} to remove all 12 columns of funds percentage gain into Funds-gain columns, shown in **Fig.2**.

```
> str(gain_annual)
'data.frame':   15 obs. of  13 variables:
 $ class: Factor w/ 16 levels "1991","1992",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ emg  : num  59.9 11.4 74.8 -7.3 -5.2 6 -11.6 -25.3 66.4 -30.6 ...
 $ sg   : num  51.2 7.8 13.4 -2.4 31 11.3 12.9 1.2 43.1 -22.4 ...
 $ mg   : num  47 8.7 11.2 -2.2 34 17.5 22.5 17.9 51.3 -11.7 ...
 $ hy   : num  46.2 15.7 17.1 -1 19.2 11.4 12.8 1.9 2.4 -5.9 ...
 $ sv   : num  41.7 29.1 23.8 -1.5 25.7 21.4 31.8 -6.5 -1.5 22.8 ...
 $ lg   : num  38.4 5.1 1.7 3.1 38.1 24 36.5 42.2 28.2 -22.1 ...
 $ mv   : num  37.9 21.7 15.6 -2.1 34.9 20.3 34.4 5.1 -0.1 19.2 ...
 $ lv   : num  22.6 10.5 18.6 -0.6 37 22 30 14.7 12.7 6.1 ...
 $ ig   : num  16 7.4 9.7 -2.9 18.5 3.6 9.7 8.7 -0.8 11.6 ...
 $ intl : num  12.1 -12.2 32.6 7.8 11.2 6 1.8 20 27 -14.2 ...
 $ tbill: num  5.7 3.6 3.1 4.2 5.7 5.3 5.2 5.1 4.7 6 ...
 $ cpi  : num  3.1 2.9 2.7 2.7 2.5 3.3 1.7 1.6 2.7 3.4 ...
> str(gain_annual2)
'data.frame':   180 obs. of  4 variables:
 $ class: Factor w/ 16 levels "1991","1992",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Funds: chr  "emg" "emg" "emg" "emg" ...
 $ gain : num  59.9 11.4 74.8 -7.3 -5.2 6 -11.6 -25.3 66.4 -30.6 ...
 $ group: chr  "Group1" "Group1" "Group1" "Group1" ...
```

Fig.2 The result after transforming the schema of dataset

2.3.2 Group Separation

To address the problem of accurate comparison and simplify appearance, we have to divide the original graph to several sub-graphs via a scientific standard. To do this, we need to observe the pattern of these 12 lines in the graph via {ggplot} package first, as shown in **Fig. 3**.

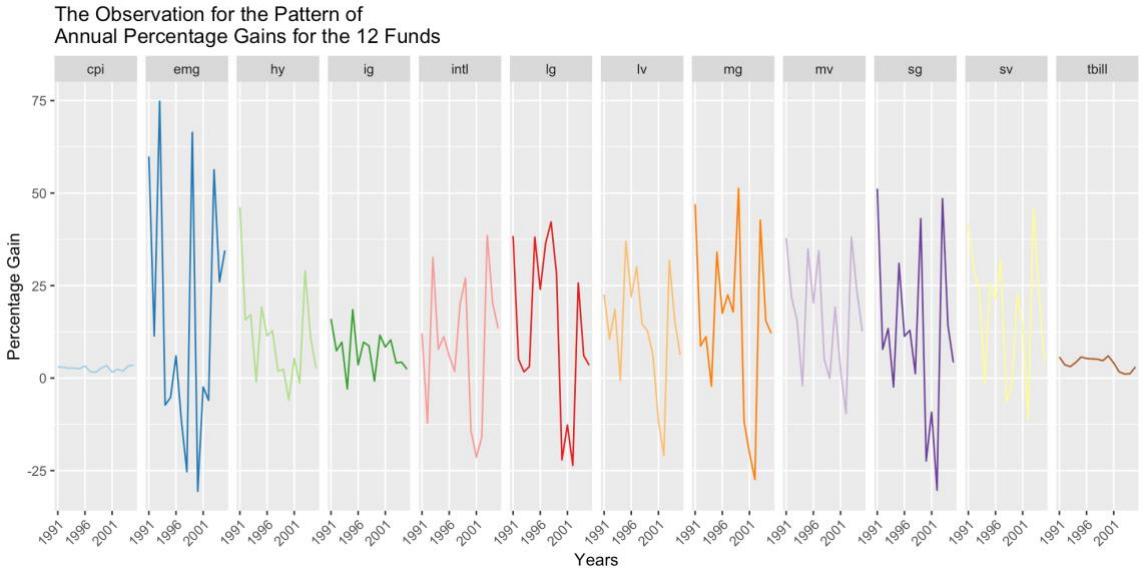


Fig. 3 The Observation for the Pattern of Annual Percentage Gains for the 12 Funds

According to **Fig. 3**, we can see that the most significant character for each fund is the fluctuation range. Specifically, “emg,” “mg,” and “sg” are more likely to have wider fluctuation range compared with the others. Also, “cpi,” “ig,” and “tbill” tend to have a steady trend. Hence, we can use the standard deviation statistic to divide all 12 funds to four groups which consist of 3 for each. Accordingly, the calculated results of the standard deviation statistic for each fund are shown in **Fig. 4**. Finally, the results for the four group are shown in **Fig. 5**.

```
> std
# A tibble: 12 x 3
    value    std funds
    <dbl>  <dbl> <chr>
 1 34.4   34.4   emg
 2 23.9   23.9   sg
 3 23.4   23.4   mg
 4 22.1   22.1   lg
 5 18.2   18.2   intl
 6 17.4   17.4   sv
 7 15.8   15.8   lv
 8 15.4   15.4   mv
 9 13.5   13.5   hy
10  5.79   5.79  ig
11  1.65   1.65  tbill
12  0.662  0.662 cpi
```

Fig. 4 Results of the Standard Deviation Statistic for each Fund

```
> (group1 <- std$funds[1:3])
[1] "emg" "sg"  "mg"
> (group2 <- std$funds[4:6])
[1] "lg"   "intl" "sv"
> (group3 <- std$funds[7:9])
[1] "lv" "mv" "hy"
> (group4 <- std$funds[10:12])
[1] "ig"    "tbill" "cpi"
```

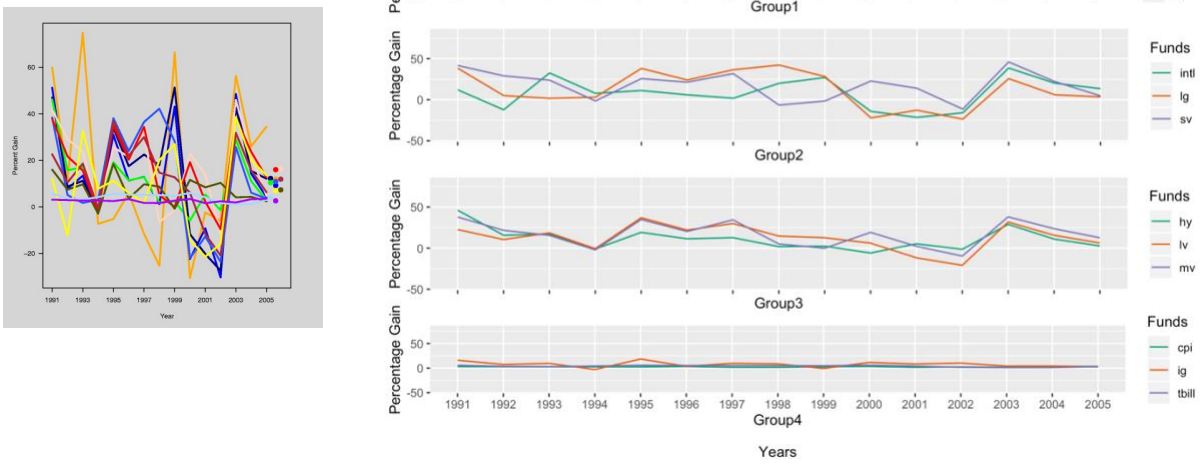
Fig. 5 Results of the final Group Separation

2.3.3 Graph Designing

Based on the separation results for the 12 funds, we use {ggplot} package to redesign the original bad graph, shown in **Fig. 6**. Specifically, not only do we divide the 12 funds via standard deviation statistic, but also add the necessary legends and title with the purpose of both good interpretation and simplify appearance. Moreover, as for the Color Brewer schemes, we choose categorical schemes to provide audience with a strong comparable appearance between each unordered element.

2.4 Redesign Results and Analysis:

Compared with the original graph, the redesigned graph improves all problems mentioned in part **2.1**.



Original Graph **Redesigned Graph**

Fig. 6 The Result for Redesign 1

- **Enable Accurate Comparisons:** First of all, compared with original graph, horizontally, we can easily observe how the trend of each fund change over time. Vertically, as for the four groups, we can easily compare the differences between each group. Specifically, from group 4 to group 1, the funds tend to have higher and higher standard deviation, and wider and wider fluctuation range. Moreover, as for the funds in each group, we can easily compare the specific details for funds with similar trend. Also, we add the grid to make it more comparable when reading the values.
- **Simplify Appearance:** After the scientific separation for 12 funds, the graph looks much cleaner and easier to understand. Each subgraph only has no more than 3 lines.

- **Interpretation:** Compared with the untitled and unlabeled original graph, the redesigned graph displays specific and concise information of the graph. We can recognize what the purpose of the graph is, what each line stands for.
- **Engage the Reader:** Simplify appearance, enough interpretations, accurate comparisons, and nice color settings make the improved graph much better to attract the audience.

3 Redesign 2

3.1 Introduction:

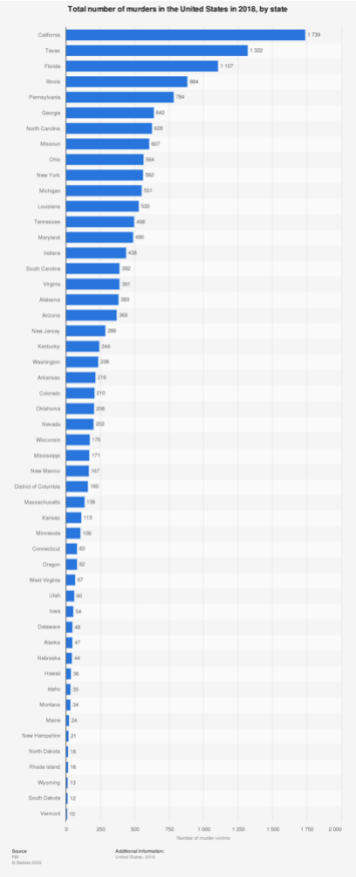


Fig. 7 Bad Graph 2

The graph in **Fig. 7** shows the Total Number of Murders in the United States in 2018 per State (Statista, 2019). According to the graph, it shows that the highest total number of Murders happens in California (Statista, 2019). However, since the number of populations in each state are not the same, general people who read the graph can assume that California is the most dangerous state by solely assuming the highest number of murders. Therefore, another variable should be considered in order to make the comparison between one state to another will look fair and make more sense. For instance, if we consider that each state owns the same population, the comparison will look better. Even more,

it also can reduce the bias as well. According to such deficiencies, as well as the first graph, this second graph will be analyzed through the following 4 aspects:

- **Enable Accurate Comparisons:** such graph can still confuse the reader since there are no boundaries between each state to another. Boundaries can help the readers if the readers want to focus in some areas of the incident. For example, if the readers want to focus on the five states which have the highest murder number and vice versa.
- **Complete the Information:** if we refer to the second bad graph, the readers only can compare the value between one state and other states in the same year. However, readers sometimes want to know how the increment or decrement compared to the year before. Since the graph does not provide any information about the position of a year before, readers can not see whether each state experiences an increment or decrement compared to the year before.
- **Interpretation:** Since this second bad graph does not fairly scale the population number in each state, the readers can misunderstand the graph and conclude that the most dangerous state is California. However, if the population numbers in each state are fairly scaled, the graph can give more reliable information to the reader.
- **Attract the Reader:** As we can see, each state on the graph only represents the same color. Also, by only using one color, the reader will find it difficult to do the comparison between one state to another. Additionally, less color can also less attract the readers.

3.2 Redesign Goals:

The purpose of presenting this second bad redesign case is to reduce the deficiencies by scaling the population of each state into the same number (per 100,000 population), giving boundaries between every five states, coloring each state with different colors, and giving the more accurate result.

After the second bad graph is redesigned, the graph should have improved the four aspects above such as providing the information more accurately comparable between one state to another, getting the information more complete than before, giving the better interpretation for the readers, and attracting the readers more than before.

3.3 Redesign Process:

The main tools used in redesigning the second bad graph is R studio, within the micromapST package.

3.3.1 Data Preparation

In order to implement the micromapST function, we should have the data scale of 100,000 population of each state and also the data of percent change rate from 2017 to 2018. Afterward, the data should be transferred to one variable and converted based on how the data will be used and performed such as shown in **Fig. 8**

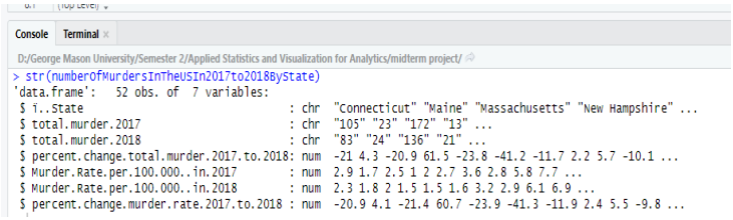


Fig. 8 The Structure of Converted Dataset

3.3.2 Graph Designing

Based on the data preparation that we already did, we use micromapST package to redesign the original bad graph, shown in **Fig. 9** In detail, the data will be separated into two parts, which are The Murder Rate Per 100,000 and The Percent Change of Murder Rate.

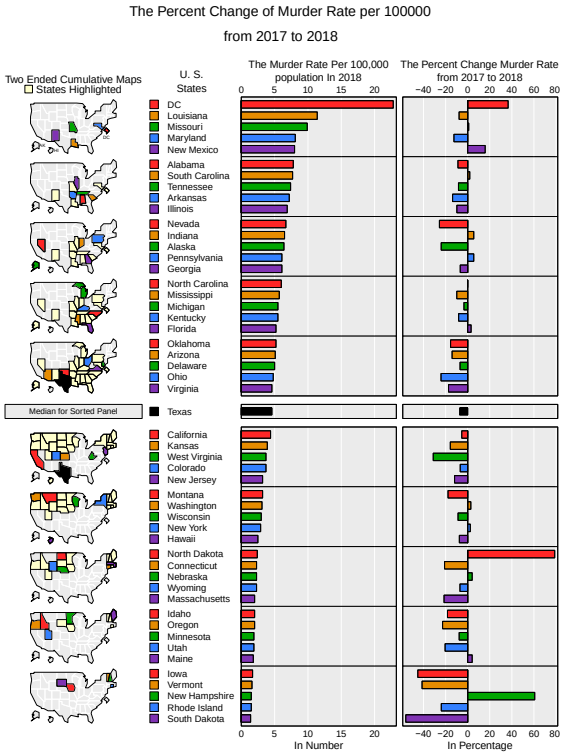
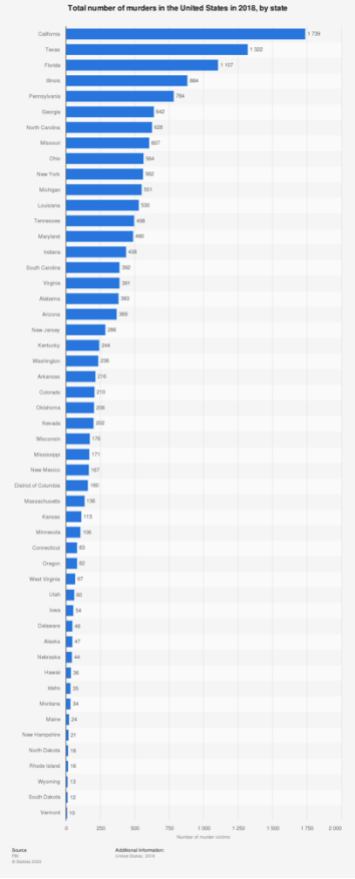
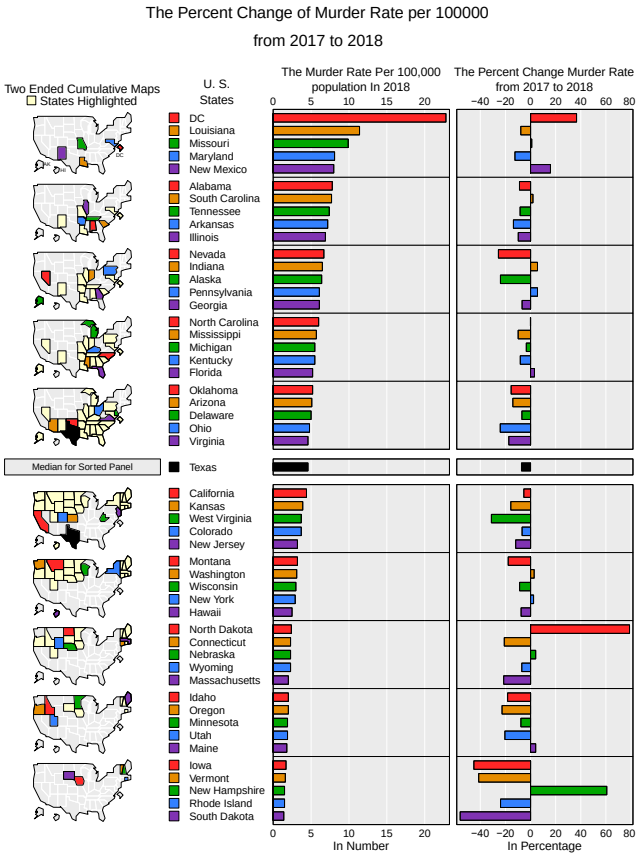


Fig. 9 Redesign Result for Bad Graph 2

3.4 Redesign Results and Analysis:

Compared with the original graph, the redesigned graph improves all problems mentioned in part **3.1**.





Original **Redesigned Graph**

Fig. 10 The Result for Redesign 2

If we refer to **Fig. 10** above, it obviously shows that the redesigned graph above is more attractive. Moreover, such graph covers the shortage of the prior graph.

Such new graph shows that the highest rate of the murder happens in DC. However, if we refer to the graph before, California has the highest number of murders. According to that, as we mentioned before, since the number of populations in each state are different, the first graph can mislead the readers since the comparison between one state to another is not fair. However, if we refer to the

second graph that we redesigned, the comparison looks fair and more reliable since the assumption of the population in each are the same.

Other than that, we assume that the second graph is more complete compared to the first graph. The second graph shows the map or the exact location of each data. Therefore, the readers can see how the pattern expands from one state to another through the second graph.

Besides that, the second graph has line boundaries between each five states. The boundaries which are shown in the second graph can ease the readers if they want to focus only on the five highest dangerous states or vice versa. Also, the colors provided in the second graph also can help the readers to see the comparison between one state to another.

Another interesting part we added in the second graph is the increment and the decrement percentage of murder in each state. As we mentioned before, since the first graph only contains data of 2018, the readers will not able to see the decrement or the increment of the murder in each state. However, if we refer to the second graph, there is additional information about the change rate compared to the year before and it can show the readers about the increase and the decrease of each state compared to the year before.

The last one, obviously, since the second graph contains more components which do not provided at the first graph, such as the map, the different color of each states, and the percent change of murder rate, the second graph should look more attractive compared to the first graph.

4 Challenges

During the process of this project, the first challenge is the selection of cases because sometimes it is hard to find the datasets for the graphs we would like to redesign. It took us almost a third time of the entire project. After that, in the first redesign case, we find it hard to assign individual legend for each subgraph when we use `facet_grid()` in {ggplot2} package. So, we design the graph for each group and then call `grid.arrange()` in {gridExtra} package to combine them.

5 Conclusion

We redesigned two different types of bad graphs in this project and improved them via four aspects, accurate comparisons, simplify appearance, interpretation, and engaging readers. During this process, we use all learned packages in STAT515, including {ggplot2}, {micromapST}, {tidyr}, and {dplyr}. Besides the experience of plotting graphs via {ggplot2} and {micromapST} packages, we learned the principles for how to make nice graphs and understand the power of {tidyr} and {dplyr}

packages for data preparation. We will explore and take full advantage of above techniques in the works and projects in the future.

6 References:

Price, P. (2006, May 23). *A bad graph but not clear how to make it better*. Retrieved from Statistical Modeling, Causal Inference, and Social Science: https://statmodeling.stat.columbia.edu/2006/05/23/post_8/

Statista. (2019, September). *Total number of murders in the United States in 2018, by state*. Retrieved from Statista: https://www.statista.com/statistics/195331/number-of-murders-in-the-us-by-state/