MID-TERM PROJECT

REDESIGNING A BAD GRAPH

Instructor:Richard Sigman

Student:Naga Sai Dhanya Veerepalli

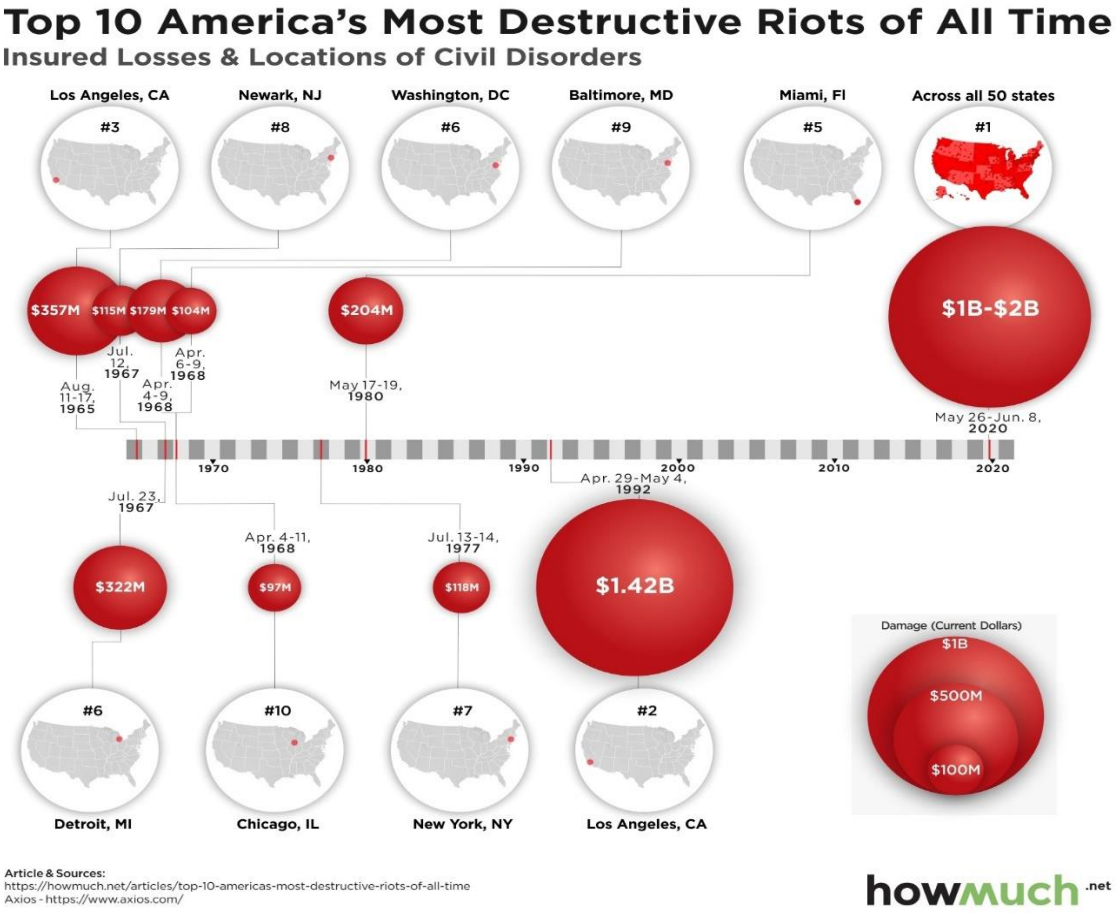
Date:19-03-2023

1.INTRODUCTION:

For the mid-term project, I have decided to redesign a bad graph into a better one using the software solutions and concepts taught in class. What are bad graphs? Bad graphs are generally challenging to analyze and make us uncomfortable inferring any information from the same. The graphs must be so that even a person without any knowledge of data visualization can read what the graph is telling. This is the main principle of (re)designing a bad graph. Some examples, if we could speak for bad graphs, can be a pie chart for a large can be very inappropriate as the legends can be confusing, An overlapping line or scatter plot, conjusted bar chart and grouped bar charts, and Stacked bar charts. In this paper, a bad graph that looks bulk and heavy for a person who needs to gain knowledge of data visualization, a bubble plot, is redesigned to make it more readable and understandable. The bad graph, its data and source, Strengths and weaknesses, the redesign objectives, techniques, and software solutions are discussed in the following sections.

2.Bad-Graph:

The bad graph discussed in this paper concerns the Top 10 most destructive riots in the USA that have caused colossal insurance losses. It is taken from the Howmuch website, which has statistical data on different topics.



There are the bad graphs chosen to be redesigned. As previously said, the above graph feels bulky and heavy to read and interpret. Here is the data that is used to visualize the above graph.

Date	Location	Insurance Loss (Current $)
Apr. 29–May 4, 1992	Los Angeles, CA	**$1.42B**
Jan. 6, 2021	Washington, DC	**$500M–$1B**
Aug. 11–17, 1965	Los Angeles, CA	**$357M**
Jul. 23, 1967	Detroit, MI	**$322M**
May 17–19, 1980	Miami, Fl	**$204M**
Apr. 4–9, 1968	Washington, DC	**$179M**
Jul. 13–14, 1977	New York, NY,	**$118M**
Jul. 12. 1967	Newark, NJ,	**$115M**
Apr. 6–9, 1968	Baltimore, MD	**$104M**
Apr. 4–11, 1968	Chicago, IL	**$97M**

The above table is based on the Top 10 destructive riots in different USA locations that caused huge Insurance losses. Riots and civil disorders can cost hundreds of millions of dollars. The longer they go on, the larger the loss is. All these riots are majorly due to political disturbances and, most notably, for civil rights. The Insurance loss is only for claimed or filed cases. Many unfiled cases can increase these figures even more if filed. The data about the insurance loss comes from the Insurance Information institute. Any figure which is over $25 million is reported as a catastrophe. Furthermore, data is adjusted for inflation.

As we can see, the data shows the riot's location and the respective insurance loss that occurred between 1965 and 2021. We can also observe that in Washington, DC, and Los Angeles, CA, the riots have occurred twice, and in Los Angeles, the riot in 1992 caused the highest Insurance loss.

Now, for this 10-line data, the graph above is complicated. Let us discuss some strengths and weaknesses of the above graph.

2.1.Strengths (of the above graph):

The above graph looks interesting as we look into it. It has location points for each place where the riot took place. These geographical map points also help one locate the place if unaware.

It also mentions the ranking of each place according to the insurance loss (higher to lower).

There are bubbles whose sizes vary according to the loss mentioned, which helps one quickly identify the location with the highest and lowest insurance loss for the occurred civil disorders.

2.2.Weaknesses (of the above graph):

Saying that again, for 10-line data, the plot looks bulk and heavy is the first observed weakness of the graph. It takes time for one person to extract the exact interpretation of the graph as the legend looks flashy and voluminous.

If we observe the graph, the dates scale has almost dates in the range of 50-55 years which looks conjusted.

The dates April 4-9, 1968, and April 6-9, 1968, almost overlap. Also, the dates July 12, 1967, and July 23, 1967, though having a gap of 12 days, overlap each other because of this conjusted dates scale.

Also, if we observe the bubbles on the left side for four states overlap. This is also due to the conjusted date scale.

3.Redesign objectives

Before redesigning a graph into a particular type, we must consider cognitive-based principles: Enable accurate comparisons, Simplify appearance, Context to improve interpretation and engage analysts or readers. If some of these principles are missing, we can consider redesigning the bad graph to improve that attribute.

The above bad graph needs to have accurate comparisons based on dates and insurance loss, simply the appearance using a simple plot or legend or label, improve the interpretation context to make it readable, and engage any person looking into the graph.

Redesigning the above graph begins with correcting the weaknesses. To begin with, it has to be decided which kind of graph would be suitable to make it understandable and easier to read. A scatterplot would be clear for small data sets in the above table.

Now to look into the plotting and grouping variables: Dates will be taken on the y-axis. The dates on the y-axis will be sorted from the latest to the oldest riots. Insurance loss corresponding to each of the locations will be on the x-axis.

Now the leftover variable location is not a continuous variable to make it a size or sequential fill attribute. It can be used as a legend with different fill colors for each location or as an in-plot label attribute. Legend attribute occupies extra space in the plot. Therefore, using a label attribute reduces the space occupied and improves clarity and data interpretation.

3.1.Software solution:

The software used to compute the above objective is user-friendly R studio. The library used to plot the graph is the ggplot2. Using this library, a scatterplot is designed using the geom_point() function. The x-axis and y-axis variables are used as the aesthetics for the ggplot function.

There are two appropriate solutions to design the above plot. One is not to fill color for the location variable, and one is to use the fill color for the location variable. In the former plot, the geom_text() function is used to text label the locations of the riots.

```
geom_text(aes(label=Location,size=NULL)
```

In the latter one, the geom_label() function is used to label the locations in the plot. Fill in the non-aesthetic attribute to give different fill colors to different locations.

```
geom_label(aes(label=Location,size=NULL)
```

These two functions were beneficial in removing the legend and saving the plot space. Now the graph will have a simple appearance, and Context for interpretation also engage a reader or an analyst, also enables comparisons.

3.2.COGNITIVE TESTING QUESTIONS AND RESPONSES:

Three cognitive testing question were asked to provide a review on the given redesined plot. The following and the questions and the responses to the same.

The responses are given based to the redesigned graphs given below.

***1**-What difference do you observe between the two graphs and which one looks better?*

***Response**- In my opinion the first plot looks better as it is more simple and clear.*

***2**-What is the first this you try to understand from the graph?*

***Response**-Insurance Loss across 10 different cities in the US between 1965 and 2021.*

***3**-Are there any properties that distract the graph from its objective?*

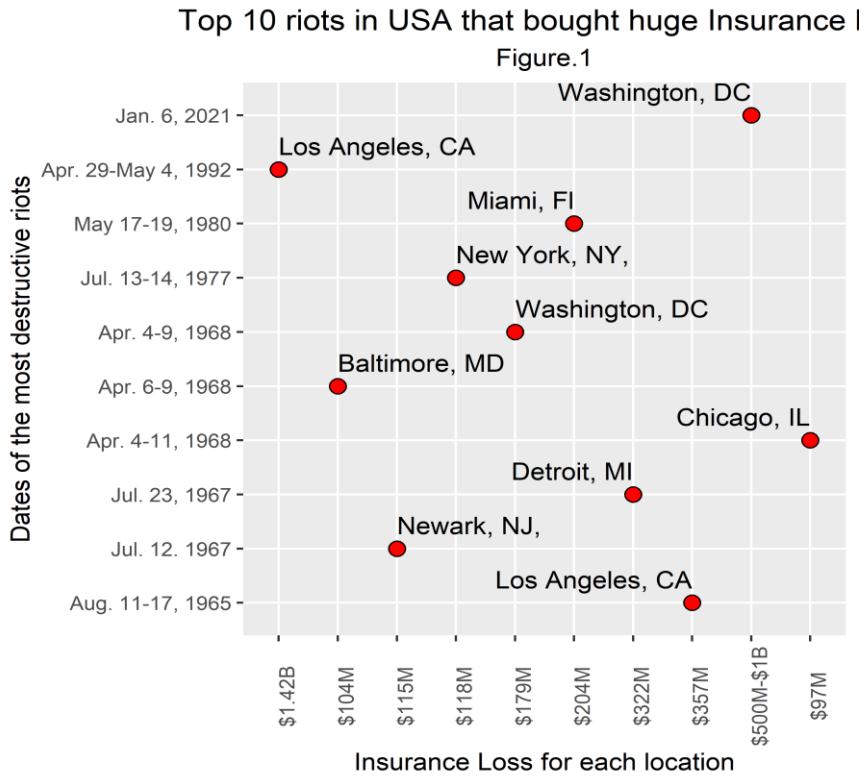
***Response**- No, the objective of the redesign is distinctly shown in the graph.*

The responses say there are no need for any changes in the graph. According to the instructor's advice the dates on the y-axis are sorted to in Descending order.

3.3.REDESIGN GRAPHS:

Redesigned graph-1:

As discussed above, the geom_text() function is used in this redesigned graph. The dates are arranged on the y-axis in descending order. This graph looks even clearer to give specific dates and insurance losses for each location. Simple text labels are given to identify the location of each point. Gridlines easily give the dates and amount of loss for each location. The background color is also in contrast to the red points. It also engages the reader in order to read the data.



Redesigned graph-2:

This plot is an improvement to the above plot. Fill color is used as a non-aesthetic attribute for the ggplot() function, which gives different colors for each place. Moreover, the geom_text() function is replaced with the geom_label() function to create text boxes and place the locations inside the boxes. We can observe that the fill color for Los Angeles and Washington also remains the same for the second time. The background color is changed from black to white to make the colors look more bright.

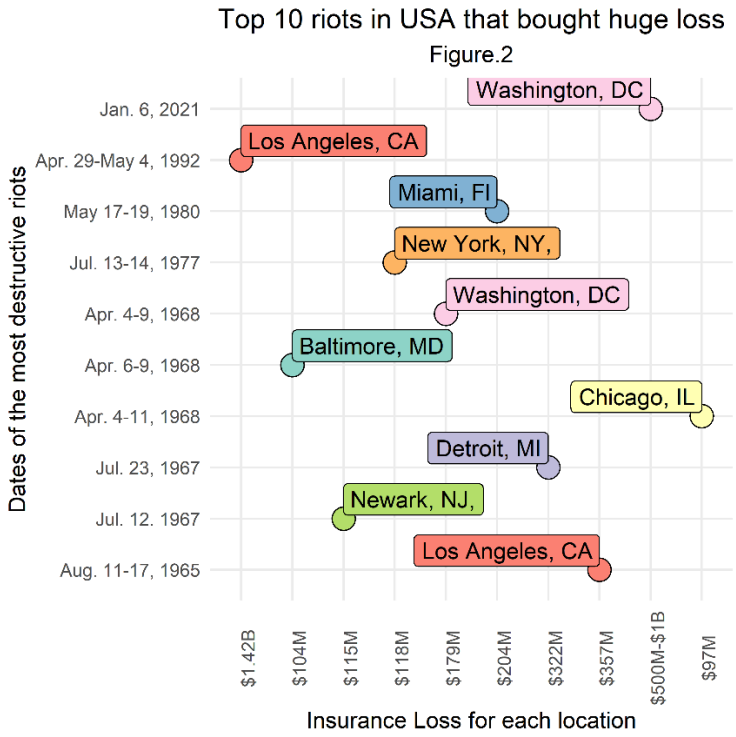


Figure.2

3.4.Challenges faced during the process:

One of the biggest challenges while redesigning the graph was choosing the type of graph that could be the best version. So, multiple attempts to create a line graph and grouped bar chart were made. However, they could have been more satisfied with reaching the requirements.

Then after realizing that a scatter or a bubble plot would give even a simple appearance for the graph, the challenge was that if a bubble plot was to be designed, the bubble size should be a continuous variable. None of the three variables in the data is a continuous variable. So then choosing a bubble plot to make it a redesign plot is rejected.

Now the only option was to create a scatterplot. At this point, the idea was simplified. The only challenge was to finalize the appearance of the plot-which variable goes to which axis and what about the third variable. Several attempts were made to give the graph an apt appearance and interpretation. Finally, the above graphs were finalized to be redesigned and satisfy all the cognitive-based principles.

4.Conclusion:

Reading a bad graph is very difficult. Hence the redesign should be an appropriate or best version of the given data set. If cognitive principles are kept in mind, and a plot is designed, data sets, whether small or large, would be the appropriate or best representation of the data. I found the redesigning process very involving and exciting.

I am looking forward to more exciting things to do with data visualization and R studio in the future.

REFERENCES:

[1] Irena.(20 April 2021). Top 10 most expensive riots in the U.S. Insurance history. Https://howmuch.Net/articles/top-10-americas-most-destructive-riots-of-all-time

[2]Hadley,Wickham.(2023/03/08).Tidyverse2.0.0.https://www.tidyverse.org/blog/2023/03/tidyverse-2-0-0/

[3]Hadley,Wickham.(2016).ggplot2. https://ggplot2.tidyverse.org/

[4]Kamil,Slowikowski.(2023-02-02).Getting started with ggrepel.https://cran.r-project.org/web/packages/ggrepel/vignettes/ggrepel.html

APPENDIX

Redesigned graphs R code

```
library(tidyverse)
```

```
library(ggrepel)
library(ggplot2)
US<-read.csv('/Users/HP/OneDrive/Documents/USA_riots.csv')
str(US)
```

```
#Redesigned graph-1
ggplot(data=US, aes(x=Insurance_Loss,y=factor(Date, level=c('Aug. 11-17,
1965','Jul. 12. 1967','Jul. 23, 1967','Apr. 4-11, 1968','Apr. 6-9,
1968','Apr. 4-9, 1968','Jul. 13-14, 1977','May 17-19, 1980','Apr. 29-May 4,
1992','Jan. 6, 2021'))))+
  geom_point(fill="red",color="black",size=3,shape=21)+
  geom_blank(data=US)+
  geom_text(aes(label=Location,size=NULL),nudge_y =
0.5,hjust="inward",vjust=0.7)+
  labs (x = "Insurance Loss for each location",
        y = "Dates of the most destructive riots",
        title ="Top 10 riots in USA that bought huge Insurance loss",
        subtitle="Figure.1")+
```

```
  theme(axis.text.x = element_text(angle = 90),legend.title =
element_text(hjust =0,vjust=0),plot.title = element_text(hjust =
0.5),plot.subtitle=element_text(hjust=0.5))
```

```
ggsave("TEXTPLOT2.png", width=7.5, height=7.5, unit="in", scale=2/3, dpi=600)
```

```
#Redesigned graph-2
ggplot(data=US, aes(x=Insurance_Loss,y=factor(Date, level=c('Aug. 11-17,
1965','Jul. 12. 1967','Jul. 23, 1967','Apr. 4-11, 1968','Apr. 6-9,
1968','Apr. 4-9, 1968','Jul. 13-14, 1977','May 17-19, 1980','Apr. 29-May 4,
1992','Jan. 6, 2021')),fill=Location))+
  geom_point(size=5,shape=21)+
  scale_fill_brewer(palette = "Set3")+
  theme_minimal()+
  geom_blank(data=US)+
  geom_label(aes(label=Location,size=NULL),nudge_y =
0.5,hjust="inward",vjust=0.7)+
  labs (x = "Insurance Loss for each location",
        y = "Dates of the most destructive riots",
        title ="Top 10 riots in USA that bought huge loss",
        subtitle="Figure.2")+
  theme(axis.text.x = element_text(angle = 90),legend.position =
'none',plot.title=element_text(hjust = 0.5),plot.subtitle =
element_text(hjust = 0.5))
```

```
ggsave("labelPLOT.png", width=7.5, height=7.5, unit="in", scale=2/3, dpi=600)
```