MTH208: Worksheet 12

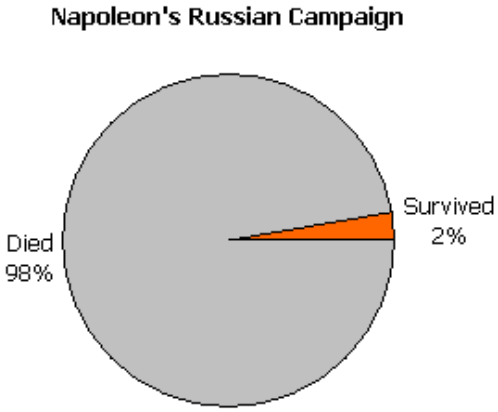
More on Visualizations

We are going to learn a little more about the importance of visualizations and telling a story. We will do this through a few case studies.

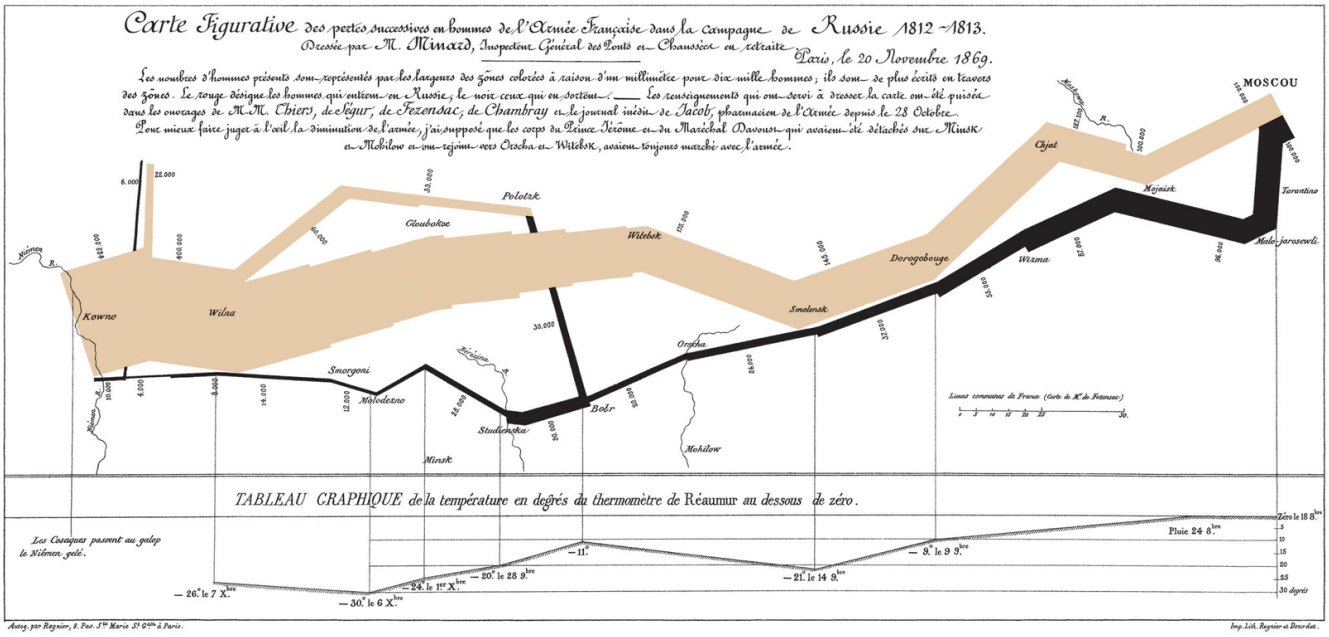
Napolean March

In June of 1812, Napolean's army of over 400,000 solders entered the Polish-Russian border and marched to Moscow to invade Russia. This is one of the most lethal military operations in history, killing about 98% of the soldiers.

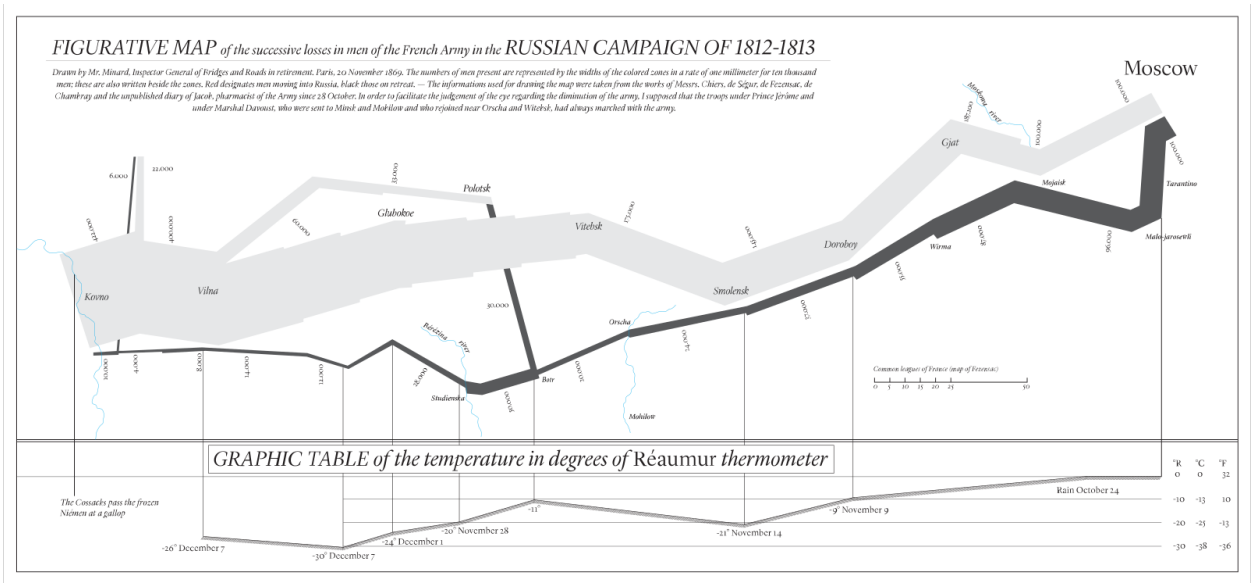
A natural visualization of the survival of the people



However, in order to tell the story of the journey and magnitude of destruction, the following visualization by Charles Minard (1869) is astoundingly informative



The following is an English translation of the map:



Impressively, the graph shows many variables in these two-dimensions:

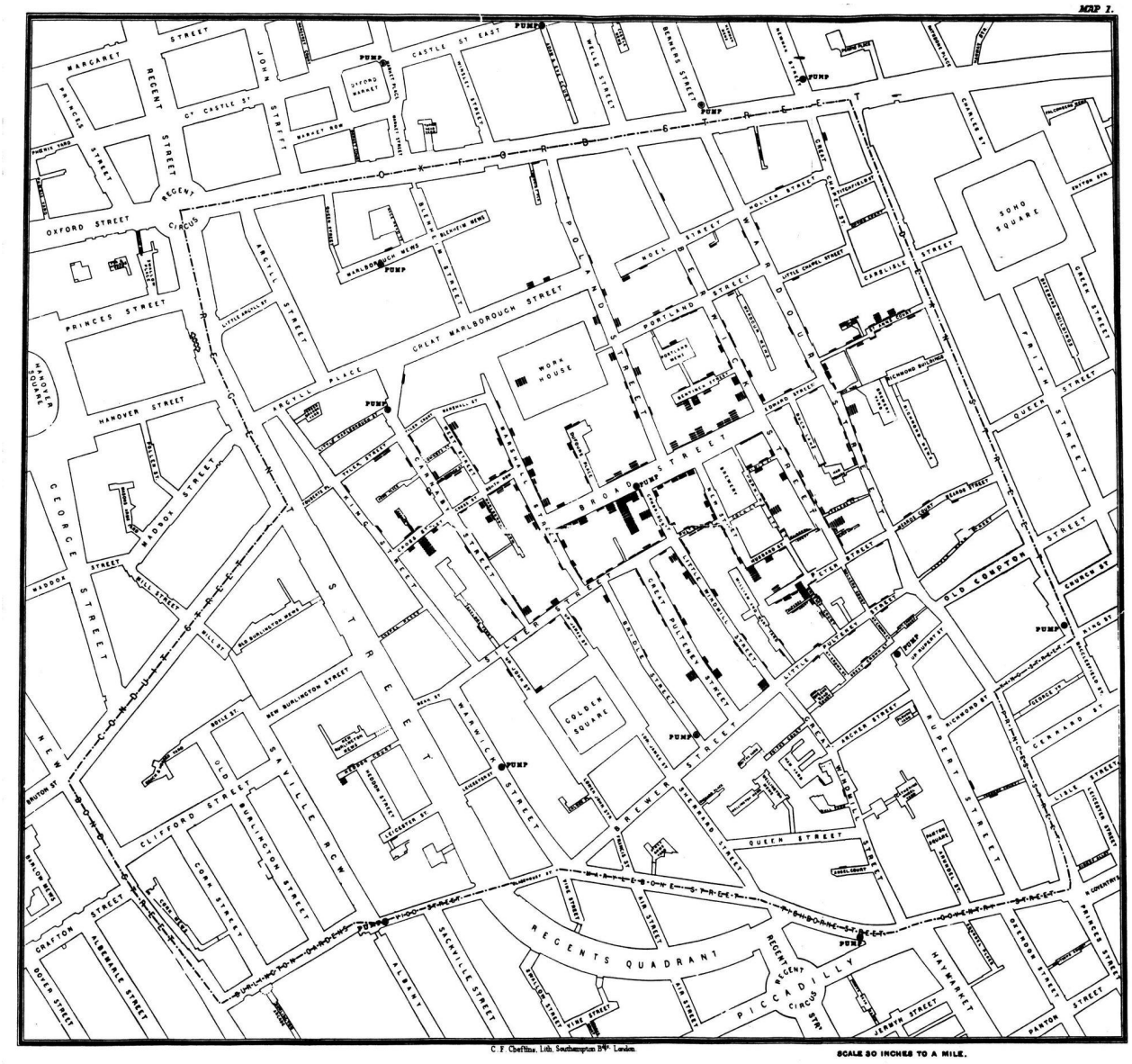
1. the number of soldiers
2. distance
3. temperature
4. direction of travel
5. local relative to dates

This plot pains us a story of the invasion and the devastation incurred.

Cholera in London

In 1854, a severe outbreak of cholera killed 616 people near Bond Street in London. At the time, not much was known on how cholera spread.

John Snow (not the Game of Throne's one) studied the cases in the area and made a modification of a *dot map* - a map with dots placed representing data.



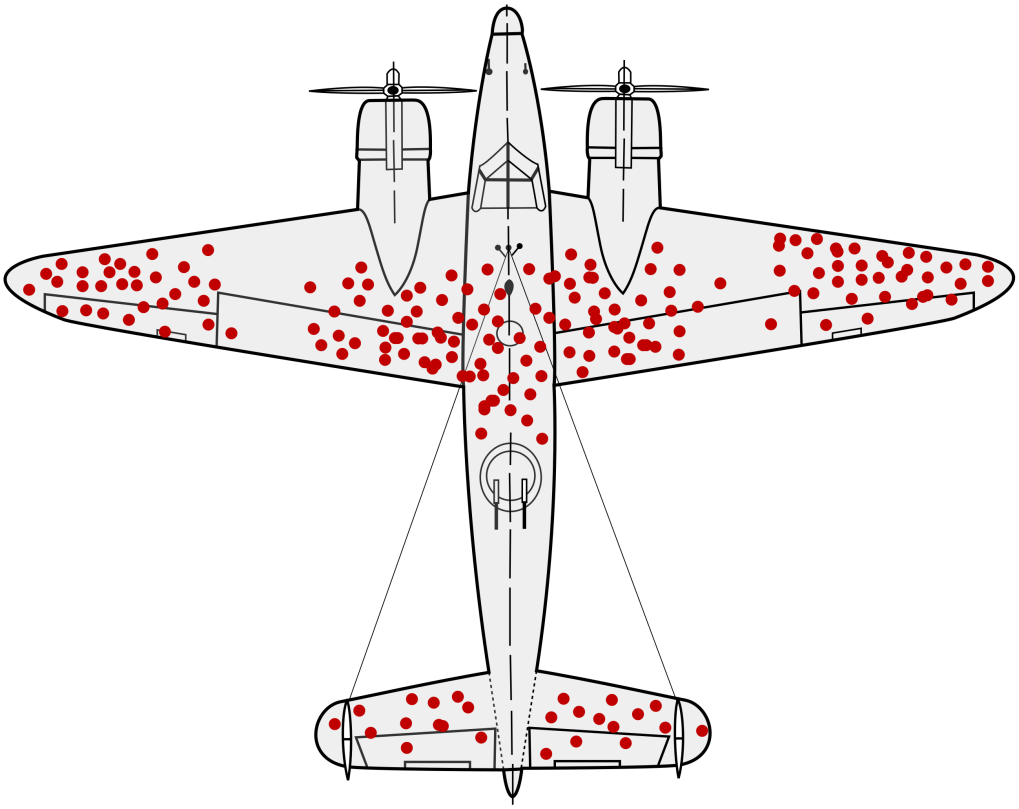
The above plot was a revelation as it was able to get to the solution of the problem. Can you guess?

Bullet Holes

Reading plots is as important as visualizing plots.

During the World War II, the US military was trying to understand how to minimize their aircraft loss in battle, and where to reinforce more armor in their aircraft. Naturally, they didn't want to put too much armor, as that increases the weight of the aircraft.

They analyzed the placement of the bullet holes on the planes that returned and arrived a map similar to the one below

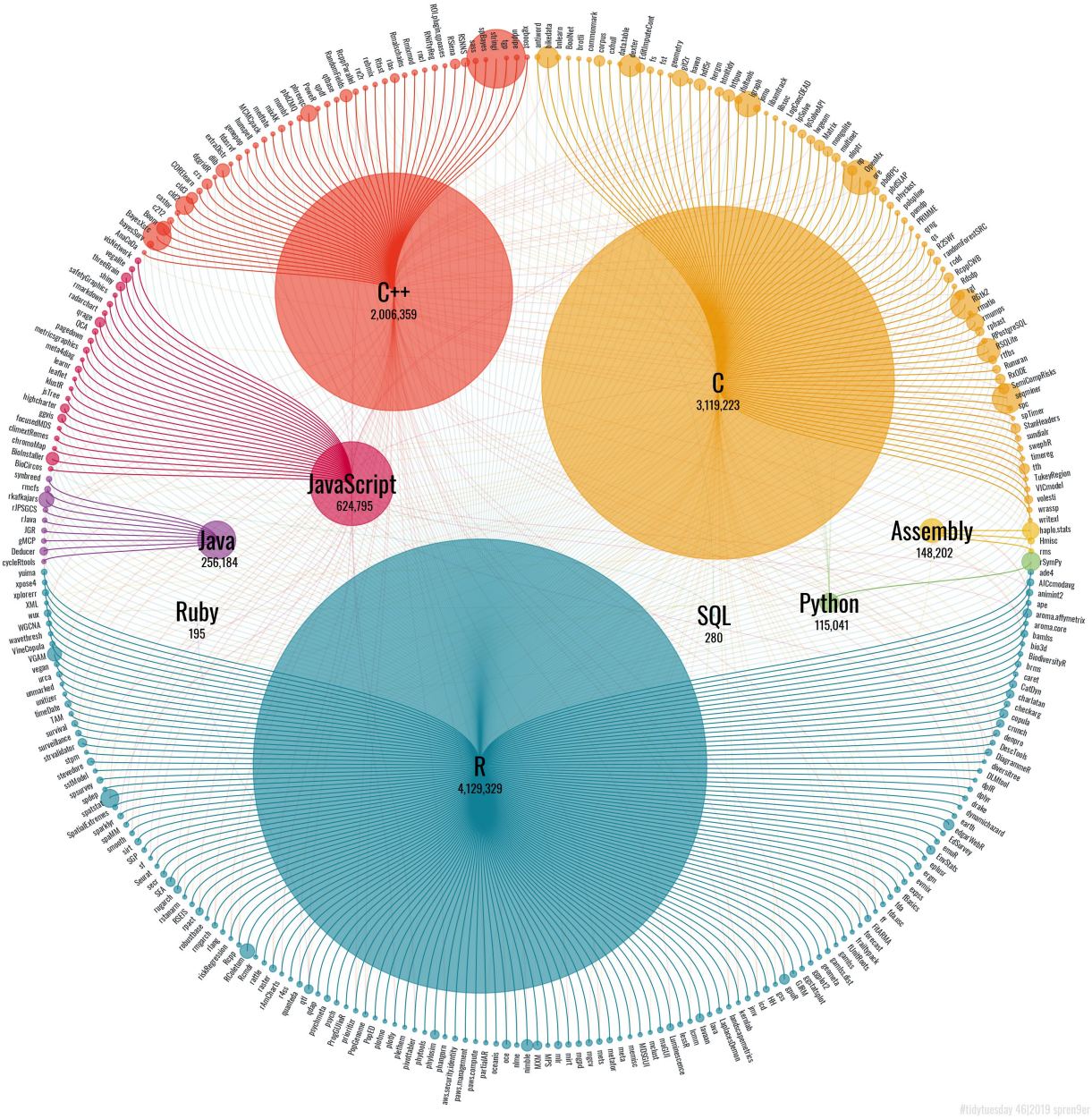


They went on to put more armor on where the red dots were found. However, this turned out to be not very helpful. Mathematician/Statistician, Abraham Wald had a different suggestion. Can you guess?

The figure of an aircraft with bullet markings is synonymous with **Survivorship Bias.**

CRAN Packages

Dr. Torsten Sprenger studied the lines of code of top 300 packages on CRAN, and produced the following visualization:



This is an example of an organized bubble/cloud plot. The code for the visualization is available here.

ggplot2

In addition to the basic R plots we learned to make in Worksheet 11, R provides another very useful graphical package for modern plots called `ggplot2`. Depending on what you want to plot and how you want to plot it, you can choose to make plots in either base R or `ggplot2`. We first load the package and load the IMDB dataset (this is in the worksheet repo).

```
library(ggplot2)
load("IMDB_movies.Rdata")
```

`ggplot2` works a little differently. It works in layers. First, we define the data are studying, then the variables, then the type of plot etc. For example, the following creates an empty plot with the right axes on the x axes.

```
ggplot(dat, aes(x = rating))
```

Below are one variable plots

```
ggplot(dat, aes(x = rating)) +
  geom_histogram()

# also run this
ggplot(dat, aes(x = rating)) +
  geom_boxplot()

ggplot(dat, aes(x = rating)) +
  geom_bar()
```

Two variable plots

```
ggplot(dat, aes(x = year, y = over.votes)) +
  geom_point()

# zooming in to some part
ggplot(dat, aes(x = year, y = over.votes)) +
  geom_point() +
  coord_cartesian(xlim = c(1996, 2025))
```

Aesthetics can be added using `aes()` options. Below I have made a plot with movies colored according to whether they released before or after 2000.

```
Year <- dat$year < 2000
Year <- as.factor(Year)
levels(Year) <- c("Before 2000", "After 2000")
ggplot(dat, aes(x = over.votes, y = rating)) +
  geom_point(aes(shape = Year, col = Year)) +
  labs(title = "Votes vs Rating", y = "Rating", x = "Number of Votes")
```

There are a number of options and features of `ggplot2()` and a lot of it can be found in their home website https://ggplot2.tidyverse.org/.

Here are some practice problems:

1. Load the `covid.Rdata` object in your repository which has updated Covid data from India. The data frame is saved in an object called `india_covid` that gets loaded when you load the `.Rdata` file. Do `names(india_covid)` to find out the names of the columns of the dataset.

2. Using examples from the website below
   http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html,

   take inspiration and make different useful visualizations for this dataset.