MATH 154 - HW2

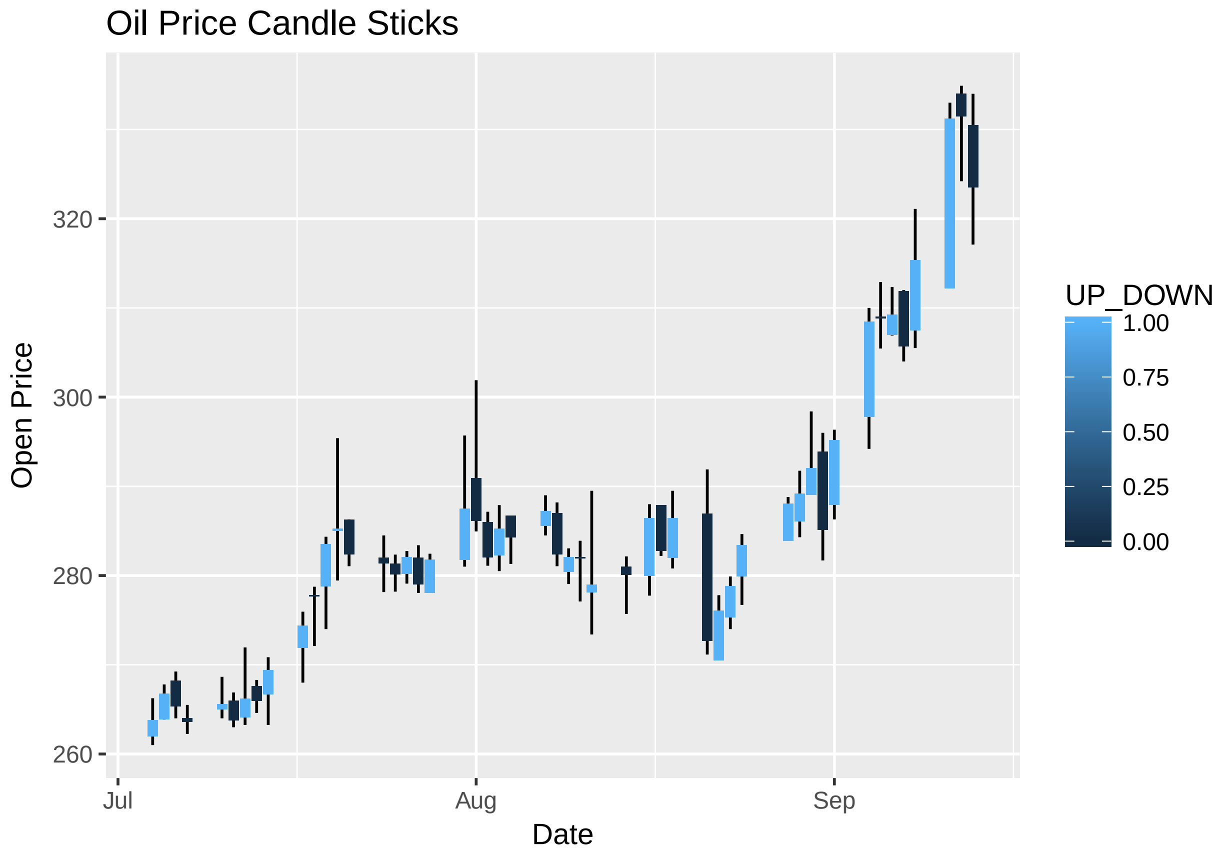
Zihao Xu

September 10, 2017

Assignment

1. Using data you find online, create a figure using ggplot (Bonus Karma Points: and shiny). Ideally, you should be able to link directly to the dataset, and not have to download the data. If you do need to download the data, make sure it is saved into your GitHub repo so that your R Markdown file can knit.

```r
library(ggplot2)
library(dplyr)
library(Quandl)

oil_price <- Quandl("NSE/OIL")[1:50, ]
oil_price$Volume <- oil_price$`Total Trade Quantity`
Width <- 1
oil_price$UP_DOWN <- ifelse(oil_price$Close > oil_price$Open,
    1, 0)

oil_price %>% ggplot(aes(x = Date)) + geom_linerange(aes(ymin = Low,
    ymax = High)) + geom_rect(aes(xmin = Date - Width/2 * 0.9,
    xmax = Date + Width/2 * 0.9, ymin = pmin(Open, Close), ymax = pmax(Open,
        Close), fill = UP_DOWN)) + labs(x = "Date", y = "Open Price",
    title = "Oil Price Candle Sticks")
```

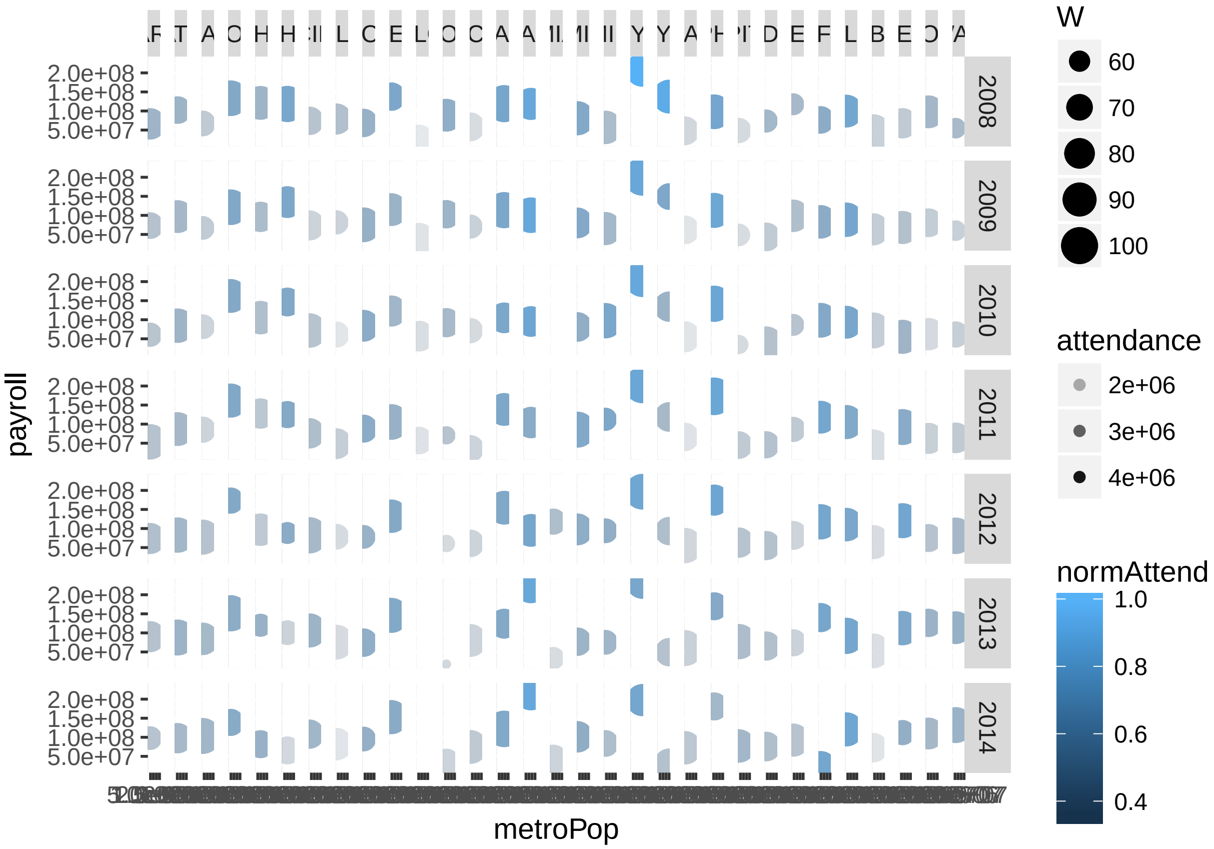


2. The `MLB_teams` data set in the `mdsr` package contains information about Major Leauge Baseball teams in the past four seasons. There are several quantitative and a few categorical variables present. See how many variables you can illustrate on a single plot in R. The current record is 7. (Taken from MDS, chapter 7.) [Note: this is *not* good graphical practice – it is merely an exercise to help you understand how to use visual cues and aesthetics!]

```
library(mdsr)
data(MLB_teams)
names(MLB_teams)
```

```
##  [1] "yearID"     "teamID"     "lgID"       "W"          "L"
##  [6] "WPct"       "attendance" "normAttend" "payroll"    "metroPop"
## [11] "name"
```

```
MLB_teams %>%
  ggplot(aes(x = metroPop, y = payroll, color = normAttend)) +
  geom_point(aes(alpha = attendance, size = W)) + # shape = name
  facet_grid(yearID ~ teamID)
```

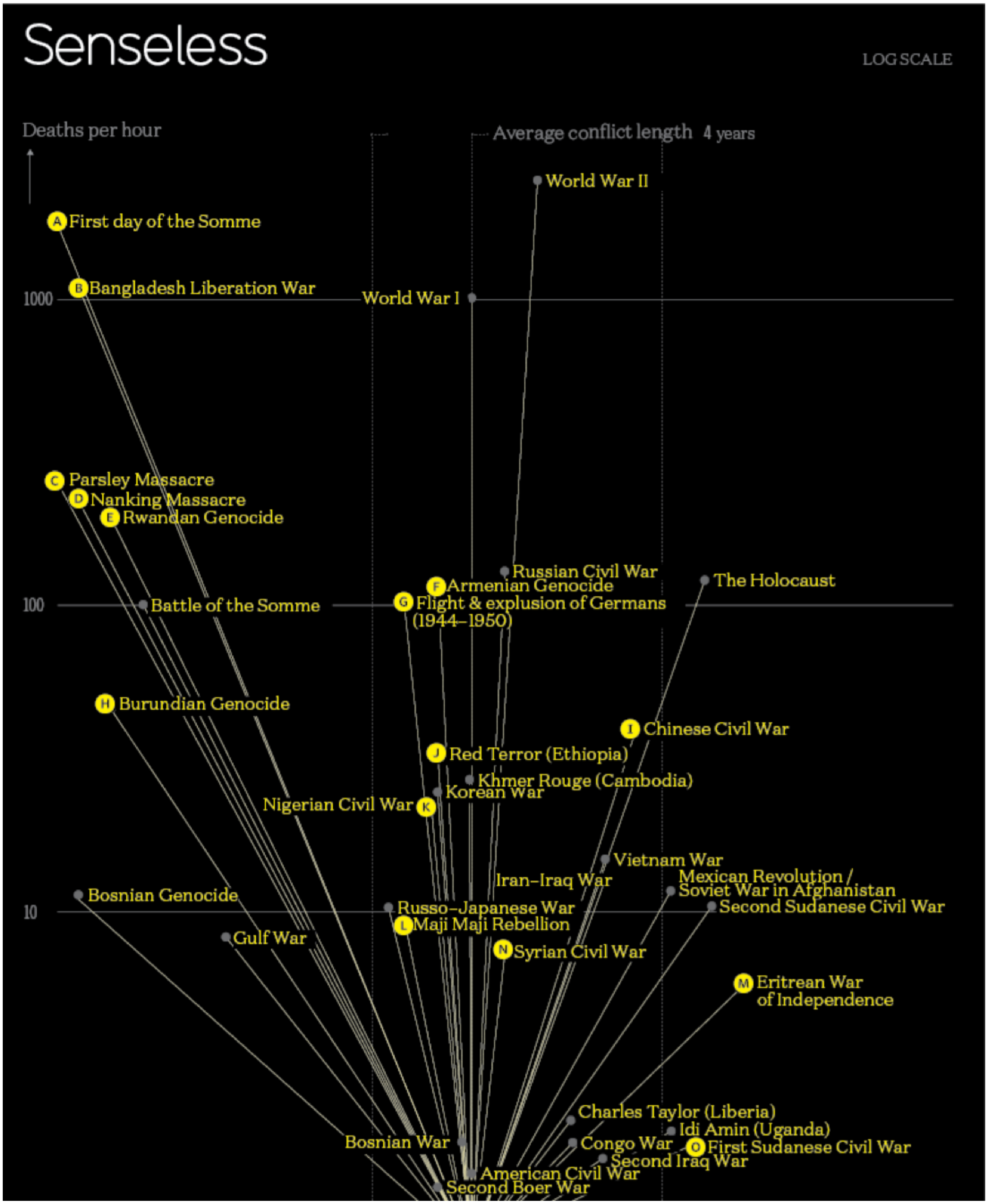


```
# Total # of variables used: 7. I am able to squeeze in 8 variables by including shape = name.
# The code complies but the plot is no longer readable...
```

3. Check out the website Information Is Beautiful http://www.informationisbeautiful.net/data/.

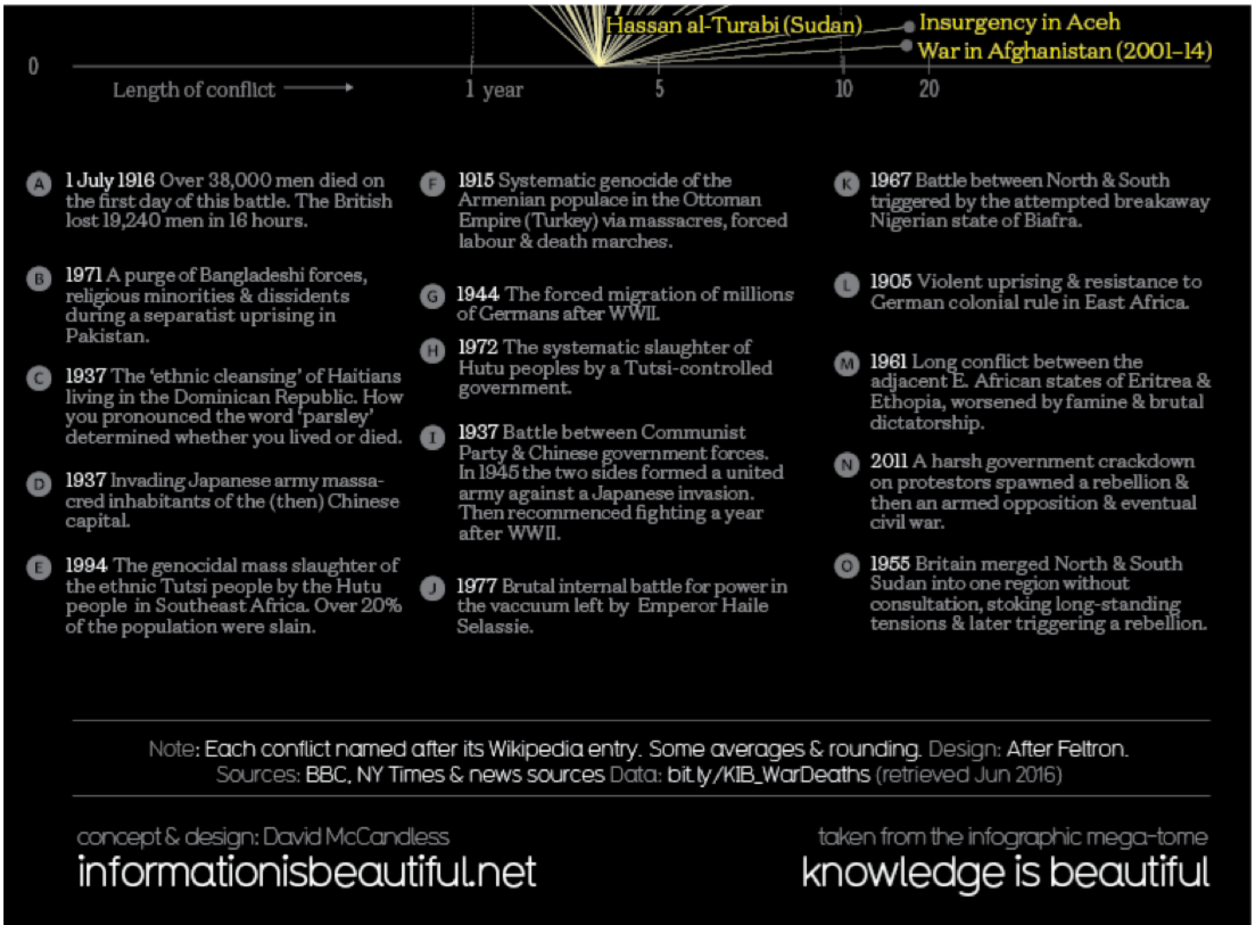
   - Find one plot on Information Is Beautiful (either save the image to your computer and upload it into your assignment (you will probably need to use `googlesheets`, https://cran.r-project.org/web/packages/googlesheets/vignettes/basic-usage.html), be sure to give the URL and citation associated with the plot; or link directly to the url for the image) that violate the concepts of effective data visualization.
   - Write a few sentences about the plot, with a critique of what aspects of the plotting could be improved. Imagine you were going to correspond with the people who designed the plot, and give them guidance about how to make a more effective depiction of the data.
   - Using the data provided from Information is Beautiful, improve the plot in some way. (You may not be able to improve the plot overall.)

(a) Picture that I am analyzing:



Senseless

LOG SCALE
Deaths per hour
Average conflict length 4 years
1000
100
10
World War II
A First day of the Somme
B Bangladesh Liberation War
World War I
C Parsley Massacre
D Nanking Massacre
E Rwandan Genocide
Russian Civil War
F Armenian Genocide
The Holocaust
Battle of the Somme
G Flight & explusion of Germans (1944–1950)
H Burundian Genocide
I Chinese Civil War
J Red Terror (Ethiopia)
Khmer Rouge (Cambodia)
Korean War
Nigerian Civil War K
Vietnam War
Iran–Iraq War
Mexican Revolution / Soviet War in Afghanistan
Bosnian Genocide
Russo-Japanese War
Second Sudanese Civil War
L Maji Maji Rebellion
Gulf War
N Syrian Civil War
M Eritrean War of Independence
Charles Taylor (Liberia)
Idi Amin (Uganda)
Bosnian War
Congo War
O First Sudanese Civil War
Second Iraq War
American Civil War
Second Boer War



1 July 1916 Over 38,000 men died on the first day of this battle. The British lost 19,240 men in 16 hours.

1971 A purge of Bangladeshi forces, religious minorities & dissidents during a separatist uprising in Pakistan.

1937 The 'ethnic cleansing' of Haitians living in the Dominican Republic. How you pronounced the word 'parsley' determined whether you lived or died.

1937 Invading Japanese army massacred inhabitants of the (then) Chinese capital.

1994 The genocidal mass slaughter of the ethnic Tutsi people by the Hutu people in Southeast Africa. Over 20% of the population were slain.

1915 Systematic genocide of the Armenian populace in the Ottoman Empire (Turkey) via massacres, forced labour & death marches.

1944 The forced migration of millions of Germans after WWII.

1972 The systematic slaughter of Hutu peoples by a Tutsi-controlled government.

1937 Battle between Communist Party & Chinese government forces. In 1945 the two sides formed a united army against a Japanese invasion. Then recommenced fighting a year after WWII.

1977 Brutal internal battle for power in the vaccuum left by Emperor Haile Selassie.

1967 Battle between North & South triggered by the attempted breakaway Nigerian state of Biafra.

1905 Violent uprising & resistance to German colonial rule in East Africa.

1961 Long conflict between the adjacent E. African states of Eritrea & Ethopia, worsened by famine & brutal dictatorship.

2011 A harsh government crackdown on protestors spawned a rebellion & then an armed opposition & eventual civil war.

1955 Britain merged North & South Sudan into one region without consultation, stoking long-standing tensions & later triggering a rebellion.

Note: Each conflict named after its Wikipedia entry. Some averages & rounding. Design: After Feltron.
Sources: BBC, NY Times & news sources Data: bit.ly/KIB_WarDeaths (retrieved Jun 2016)

concept & design: David McCandless
informationisbeautiful.net

taken from the infographic mega-tome
knowledge is beautiful

(b) Comments on the graph Problems:

- The x and y axis of this graph do not relate that much to each other in that, the span of the war has no direct relationship with death per hour. Such combination does not show a trend as time changes nor convey too useful information;
- The lines stemming from (4,0) seems visually impressive but distracting. The origin of (4,0) might represent the "center" of x-axis when it is in log scale, but such arrangement does not add to more information;
- It is hard to contrast or compare one war to the other in terms of size: both size of conflict (domestic or international) and total number of death (the product of x, y and # of hours per year).

My suggestions:

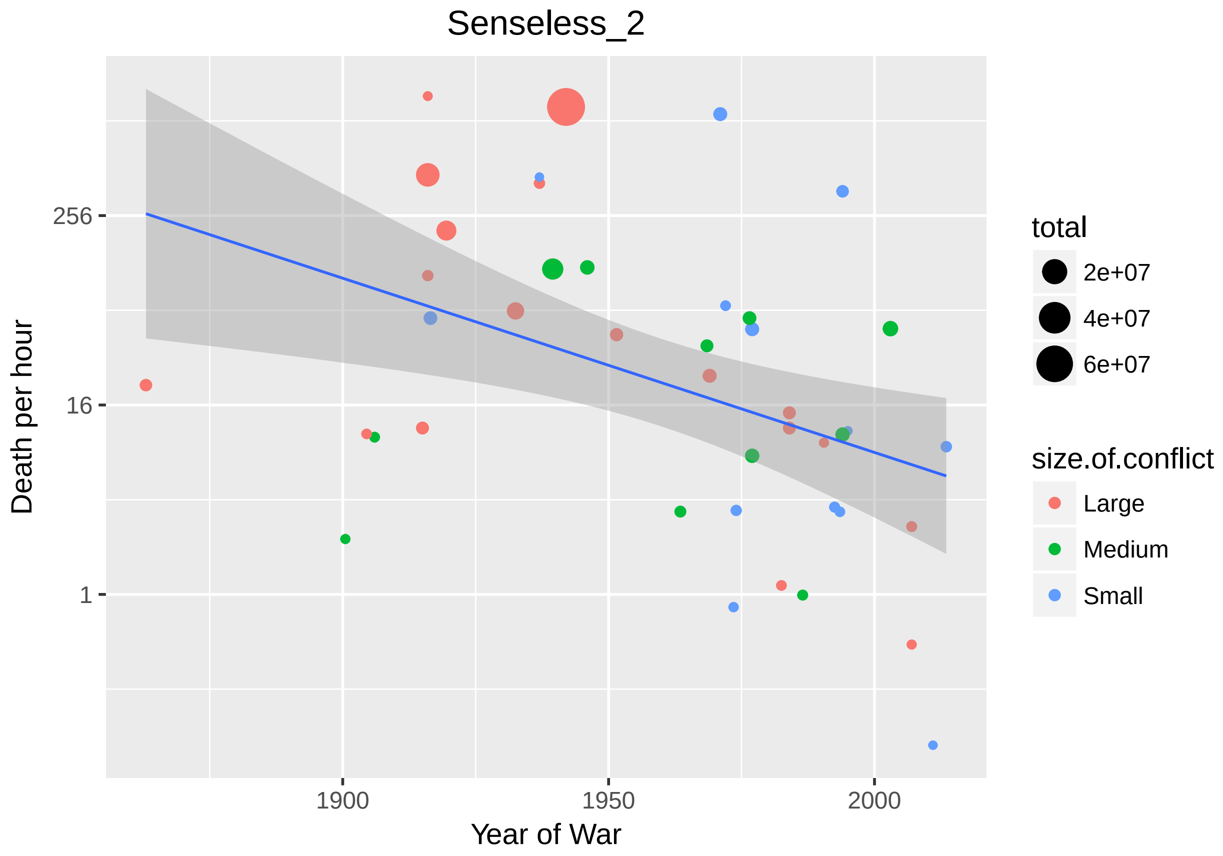
- Change the x axis to be something more informative, such as the starting time of the war;
- Delete the distracting lines orginating from (4,0);
- Use another variable, such as the total death of war, to represent the size of war, and reflect the sizes/color through sizes of the points.

(c) The plot I made:

```r
library(ggplot2)
library(googlesheets)
library(tidyr)
library(dplyr)
library(httpuv)
library(stringr)
library(forcats)
```

```r
# Citation information: David McCandless, v2.0, Jul 2016,
# Research: Miriam Quick, Interaction Design & Code: Fabio
# Bergamaschi, URL:
# http://www.informationisbeautiful.net/visualizations/senseless-conflict-deaths-per-hour/
# the dataset is obtained from the following link:
# https://docs.google.com/spreadsheets/d/1q3UnBwPgo_HsWuL8e4RFgPSDGlH7GZwiCNfv6805OkY/edit#gid=2

# Read in the csv file
senseless <- read.csv("Senseless_Conflict_Datasheet.csv")

# Get rid of the unnecessary columns
senseless <- senseless[1:42, !(names(senseless) %in% c("X", "X.1",
    "X.2", "X.3", "how.many.years.", "per.hour.1"))]

# Data cleaning
senseless$total <- as.numeric(str_replace_all(as.character(senseless$total),
    ",", ""))
senseless$per.hour <- as.numeric(str_replace_all(as.character(senseless$per.hour),
    ",", ""))
senseless$war.year <- (senseless$start.year + senseless$end.year)/2

# Refectering 'size.of.conflict' to small, medium and large
senseless$size.of.conflict <- factor(senseless$size.of.conflict,
    levels(senseless$size.of.conflict)[c(3:10, 2, 1)])
senseless$size.of.conflict <- fct_collapse(senseless$size.of.conflict,
    Small = c("small country", "small region", "civil war"),
    Medium = c("large region", "medium country", "medium region"),
    Large = c("global", "international", "large country"))


# Plotting war.year against per.hour
senseless %>% ggplot(aes(x = war.year, y = per.hour)) + geom_point(aes(size = total,
    color = size.of.conflict)) + geom_smooth(method = "lm", lwd = 0.5,
    degree = 0, span = 2/3) + scale_y_continuous(trans = "log2",
    name = "Death per hour") + scale_x_continuous(name = "Year of War") +
    ggtitle("Senseless_2") + theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: Ignoring unknown parameters: degree
```



My plot: - Uses the year of war (defined by the middle year of entire duration the war) as the x-axis, showing time progress; - I colored the points by the size of war (the refactored "size.of.conflict") and make their sizes propotional to their total death to represent severity; - I added a smooth line to represent the general decreasing trend in death per hour as time progress: I guess this is due to advancement in military technologies; - Overall my plot is more informative: showing the sizes and scale of each war, while also revealing the general decreasing trend in death per hour.

4. Describe (at least) 4 substantial ways that the poster winner "Congestion in the sky" (from the Data Expo 2009 poster competition results, http://stat-computing.org/dataexpo/2009/posters/) could be improved, using the concepts of effective data visualization. Write a constructive criticism that gives suggestions for improvement on each aspect that you criticize. (Note that a poster is different from one image in a paper or talk.)

My critique:

1 - The biggest problem with this poster (in my opinion) is the disconnection between plots and the information. The information that each plot is conveying is listed under "Goal" section to the left, while all the plots are all displayed on the right. Such indirect layout is hard for the reader to connect the useful information on the left with its respective graph on the right (they might even have to make guesses). My suggestion is to put the message next to its corresponding graph (or in its corresponding section) so that the connections between plots and messages can be more easily drawn;

2 - The "overview" section seems to be a little bit problematic in that I find discrepancies between its goal (listed on the left) and what it is actually convening (by the plots). According to the goals, this section is dedicated to "Summarize data by time periods, airport, and carrier". The current graph seems to be plotting the percentage of delay and percentage of canceled flights for each day of the year, for each airport. So I believe this design is displaying the distribution of delayed/canceled flights for different destinations and times, instead of summary statistics of the mentioned variable. Therefore, if the author's intent was to really describe the data, I would use scatter/box/histogram to describe the distribution/feature of a single or a

combination of variables; otherwise, I would change the goal of the overview sections to: an overview of the distribution of delay and canceled flights;

3 - A common problem with all the plots in this poster is that the font sizes for the title, the labels and the label tickers are too small. It is hard for the reader to understand the content nor the message if they are too small. I would suggest adjust the tittle size to be as large as needed while also increasing the sizes of the labels and tickers (but slightly smaller than the title);

4 - The choice of color and size of each section can be better fine-tuned so that some of the plot do not overshadow others (when the amount of information each conveys are basically the same). For example, the plots in the "overview" section and the major plot in the "carrier effects" section seem to have very bright coloring and large size to the degree that other plots in the same/other sections seems "less important". I would probably change the color mapping in the "carrier effects" section to be consistent with the color used in the "overview" section, and increase the size of "temporal effects" section and eliminate some of the less informative (and too small) graph in the "spatial effects" and "carrier effects" sections.