For this discussion I decided to tackle topic 2. After reviewing the handout, I think the best visualization illustrated in that article has to be the one from FiveThirtyEight, “Who Gets Paid Time Off?”. Not all that surprising since this is a site for data scientists, statisticians, and mathematicians. According to the criteria outlined by the author in the handout, this visual covers all aspects of ACCENT.

The visual is easily perceived and is comprehensible, one can distinguish the different elements, there is a consistency to the design, portrays complex data easily, there is no extras on the visual that distract from its main message, and shows an accurate picture of the data.

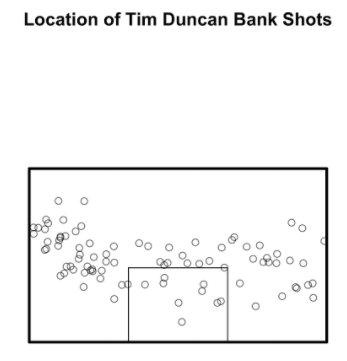
Easily the worst visual in the handout is from Cha rtjunk. Based on the principles outlined in the article, this visual does not meet the standards for any of the categories.

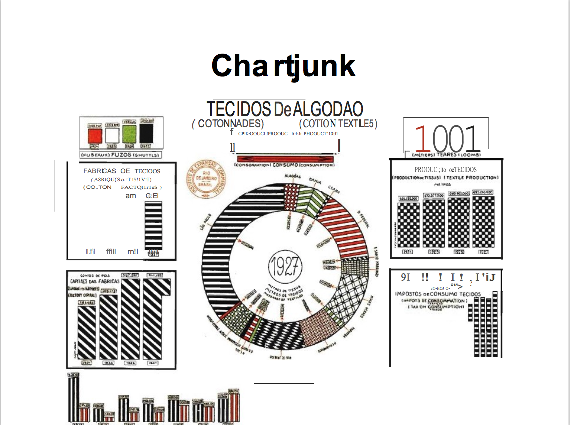
* The variables are difficult to perceive, no clear understanding of the story the visual is trying to tell
* Cannot distinguish the elements, patterns do not help visualize
* Does not portray data in a simple way, maybe makes it more complicated
* The graph in the middle does not need to be a pie chart, better to create a bar chart to display this, without the patterns
* Inability to determine the value of the graphs.

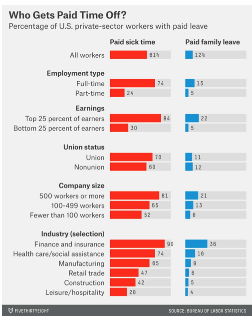
When looking at visualizations of data, it all depends on the story the analyst is trying to tell. I’m a big believer in tailoring your visual based the story and finding a creative way to illustrate this. One of my favorite examples is the visual on inpredictable1, which showcases the location of where the ball strikes the backboard whenever Tim Duncan shoots a bank shot. For those unfamiliar, a bank shot is when the backboard is used to make a shot.

The author could’ve easily just displayed a bar chart that showed how many times the ball hit the upper left quadrant of the ball, the lower left quadrant, upper right, and lower right. However, Beuoy decided to take a different approach to better let the reader visualize where the ball was hitting the board. Although a bit unconventional and not all of the ACCENT methodology is met here, I still believe the purpose of the visual was well served. Looking to see what you all think in this discussion.

1. Beuoy , M. (2016, July 16). Bonus Tim Duncan Chart - Bank Shots. Retrieved April 6, 2017, from <http://www.inpredictable.com/2016/07/bonus-tim-duncan-chart-bank-shots.html>







Discussion Weekly Problem:

I decided to take another go at one of the analyzing databases questions at the end of chapter 3. I went ahead and solved for number 5. The R code is listed below. Based on the plot there is a wide range of births at the hospitals. Which is the case in real life, not every hospital in the world is going to have the same amount of patients and will not have the same amount of births. That being said, the hospital with the largest amount of patients also had the largest amount of births. There seems to be some correlation here, but further analysis is required.

If you are replicating the R code below, make sure to change the file path and name of the csv file to your machine and preference.

#Discussion Week 2 - Chapter 3 Number 5 - Analyzing the Databases

#read file

Hosp\_DB <- read.csv(file = "/Users/Zeeshan/Desktop/PREDICT 401/Hospital.csv", header = TRUE, stringsAsFactors = FALSE)

#check structure of file

str(Hosp\_DB)

#create window

par(mfrow = c(1,1))

#create plot

boxplot(Hosp\_DB$Births, data = Hosp\_DB, main = "Births at Hospitals", ylab = "Number of Births")

Discussion Comment:

Sean, I would agree with you that FiveThirtyEight in general has the best visualizations. For my discussion I also used a visual that was created by one of the authors on that site. I took a similar approach to your visual on Cholera in 1854. I think presenting the data in a way that will help tell the story to the reader is the best possible way to get one’s point across. I used a similar visual that utilized a backboard to show where the ball hits the glass when Tim Duncan shoots a bank shot. Using a visual map as a tool can be very helpful. However, I disagree that this map is the best, I think it could have been better. There should be some color to indicate larger concentration instead of just using a circle. I’m not sure when this was created, so there is that as well, also some of the resolution could have been damaged as part of the copy/pasting. Also I find that maps that have a distance scale don’t really provide any quantitative value, but I think it’s good to help the reader get a general idea of what the distance between the Cholera occurrences are.

Discussion 2 Comment:

I tried to get a ggplot box plot, for the same problem. However, it kept asking me for a x scale and I’m a bit confused as to why. If anyone knows what’s wrong, I would appreciate the help.

#Discussion Week 2 - Chapter 3 Number 5 - Analyzing the Databases

#read file

Hosp\_DB <- read.csv(file = "/Users/Zeeshan/Desktop/PREDICT 401/Week 2/Hospital.csv", header = TRUE, stringsAsFactors = FALSE)

#check structure of file

str(Hosp\_DB)

#create window

par(mfrow = c(1,1))

#create plot

boxplot(Hosp\_DB$Births, data = Hosp\_DB, main = "Births at Hospitals", ylab = "Number of Births")

#create plot with ggplot

library(ggplot2)

births <- data.frame(Hosp\_DB$Births)

myColor <- "#FF6347"

birthbox <- ggplot(births) +

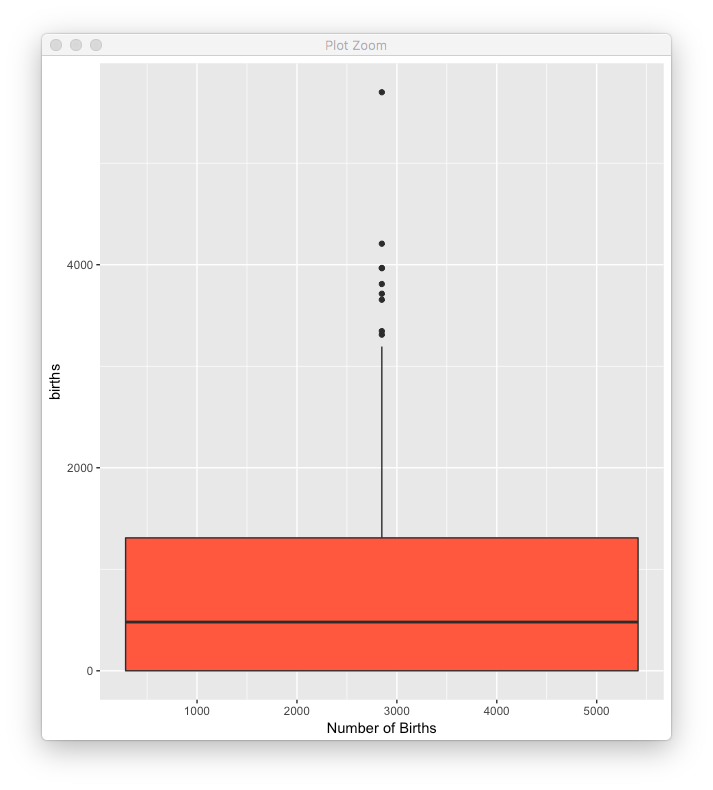
aes(x = births, y = births) +

geom\_boxplot(fill = myColor) +

xlab("Number of Births")

hist + ylim(0,6000)

print(birthbox)



Discussion 2 Comment 2 – new problem

Analyzing DB number 1

I decided to tackle another problem that involves some data analysis for practice. In chapter 3, Analyzing Databases, number 1. After pulling the data from the manufacturing DB and looking at the column for New Capital Expenditures, the mean is significantly lower than the median. The mean is about half the value of the median. This tells us that the data is heavily skewed right and that there are several outliers that are bringing up the value of the mean.

Median: 483.0

Mean: 995.2

