Economics 1660: Big Data Prof. Daniel Björkegren

PS 2: Visualization

1. Graphing Napoleon's March, Continued (optional)[1]

Revisit the graphs we produced in PS 1. The grammar of graphics package (ggplot2) makes it easy to layer different dimensions onto the same plot. For example, the following layers the location of the armies, direction of the march, the number of survivors, and city names.

Reproduce Minard's plot with the following code. I recommend adding each ggplot command one at a time so see the effect on the plot. Include your plot in your writeup. Feel free to alter the plot to make it more legible or informative.

```
library(ggplot2)
library(scales)

troops <- read.table("tables/minard-troops.txt", header=T)
cities <- read.table("tables/minard-cities.txt", header=T)

plot_minard <- ggplot(troops, aes(long, lat)) +
  geom_path(aes(size = survivors, colour = direction, group = group),
linejoin = "round", lineend = "round") +
  geom_text(aes(label = city), size = 4, data = cities) +
  scale_size(range = c(1, 12),
    breaks = c(1, 2, 3) * 10^5, labels = comma(c(1, 2, 3) * 10^5)) +
  scale_colour_manual(values = c("grey50","red"))

ggsave(plot_minard, file = "minard.pdf", width=12, height=3)
```

(Also see http://motioninsocial.com/tufte/ for examples of implementing various Tufte plots in R.)

[1] Source: Hadley Wickham

2. Top Pop[2]

How has popular music changed over time? Billboard maintains a list of top singles, the hot 100. We've compiled a dataset of these top songs. This part of the assignment allows you use this data to visualize trends in popular music.

You can download billboard.csv from the course website, and open it in R using the command:

```
billboard = read.csv("tables/billboard.csv")
```

Part A
First, let's understand the data:
- What years does the data cover?
- What is the difference between week_peak_position and overall_peak_position?
- What song has been in the charts for the longest number of weeks?
- What is the top song in the most recent chart?

Part B
Next, let's explore patterns in the data using visualizations.
I recommend reading the following blog post:
http://www.modestinsights.com/analyzing-the-billboard-hot-100/
which analyzes the chart trajectory of different songs, the Beatle's popularity in 1964, and patterns in artists' careers.

(i) Replicate the plot of the Beatles' popularity in 1964 with the following code:

```
library(ggplot2)

billboard$week = as.Date(billboard$week, format = "%m/%d/%Y")
billboard$entry_date = as.Date(billboard$entry_date, format = "%m/%d/%Y")

# Beatles Dominance in 1964
beatles = subset(billboard, billboard$artist == "BEATLES" & billboard$week
>= as.Date("1964-02-01") & billboard$week <= as.Date("1964-06-01"))

ggplot(beatles, aes(week, this_week_position, group = song, col = song)) +
scale_y_reverse() + geom_point() + geom_line() + ggtitle("Beatles
Dominance in 1964") + ylab("Position")
```

This plot shows time on the x-axis and position on the y-axis. But we could represent the

[2] Adapted from Hadley Wickham; licensed under the CC Attribution-Noncommercial-Share Alike 3.0 License. Borrows from http://www.modestinsights.com/analyzing-the-billboard-hot-100/

information in a different way. Instead of putting rank on the y-axis, we could display the track name and use another aesthetic to display the rank.

- List possible aesthetics (other than position) that we could use to display rank, and then create plots using those aesthetics.
- What do you notice? What aesthetic makes it easiest to see the pattern over time? What makes it hardest? Does combining aesthetics make it better? What makes the original plot so good?
- You've probably used size as one of the aesthetics - but what is size really mapped to? Create a small experiment to determine whether size is mapped to radius or area? Which do you think it should be mapped to?

(ii) Explore

Use the data to explore questions that you find interesting. We encourage you to choose your own investigations; some ideas:

- How has the popularity of female artists changed over time?
- How have pop musician career trajectories changed over time? Do artists peak sooner or later, and is their popularity more or less stable?
- How did pop music evolve in the 1990s?

Use ggplot to produce at least 2 informative plots for your investigation. For each visualization you produce, produce two alternates that present the same information using different aesthetics. Describe which visualization you find most informative and why.

Summarize what you've learned in a half page writeup that refers to your visualization.

3. Poisonous Mushrooms

A foodie friend wants to cook a dish with fresh collected mushrooms. However, he knows that some wild mushrooms are delicious and others can be deadly.

'There is no test to determine edible versus poisonous mushrooms. Ignore any advice such as "a poisonous mushroom will tarnish a silver spoon," "if it bruises blue, it's poisonous," etc. These are old wives' tales and folk myths, and completely untrue.'[3]

You've managed to collect data on 8,124 mushrooms, their features, and whether they are edible. It would be difficult to remember all of these individual mushrooms, so your goal in this assignment is to determine if there are rules of thumb that can help your friend.

The data is in the file "agaricus-lepiota.data.txt"

First, let's graph the data to try to identify patterns using the skills we've learned so far. The data records 22 different features for each mushroom, so we have a high dimensional space.

Explore different graphs of the data trying to identify which dimensions appear to be important for predicting whether a mushroom is edible. Include the dimension 'edible'.

You can represent many dimensions, using the x-axis, y-axis, colors, shapes, sizes, and small multiples (facets).
Hint: you can use geom_jitter, or set the opacity between zero and 1 (alpha) to prevent overplotting.

Explore the data to come up with a rule of thumb to help your friend identify edible mushrooms. Write your rule of thumb and include graphs that demonstrate its performance.

[3] http://mdc.mo.gov/discover-nature/outdoor-recreation/mushrooming/basic-mushrooming