479 HW 1

Caitlin Bolz

2/3/2021

```
library("readr")
library("ggplot2")
library(dplyr)
library(ggrepel)
```

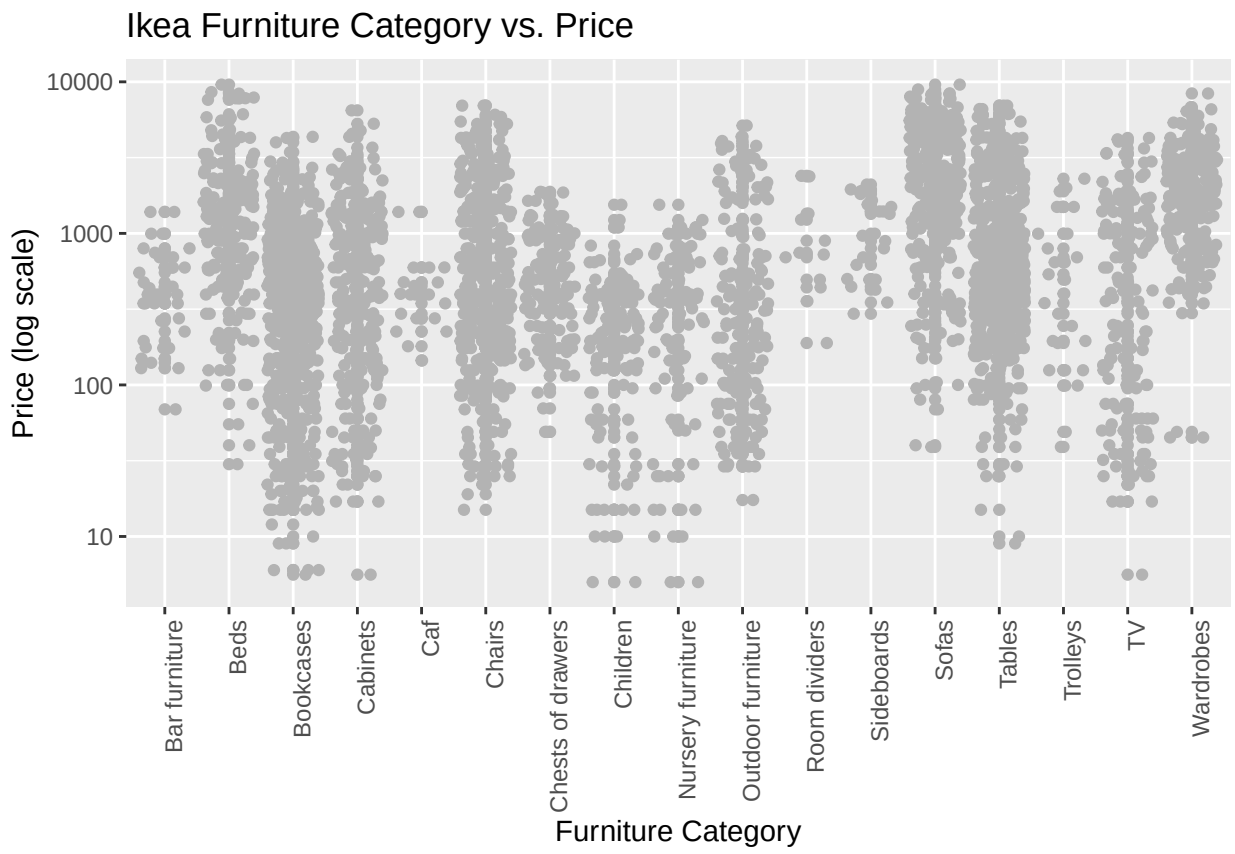
Problems

(1) Ikea Furniture

```
ikea <- read_csv("https://uwmadison.box.com/shared/static/iat31h1wjg7abhd2889cput7k264bdzd.csv")
```

Part A

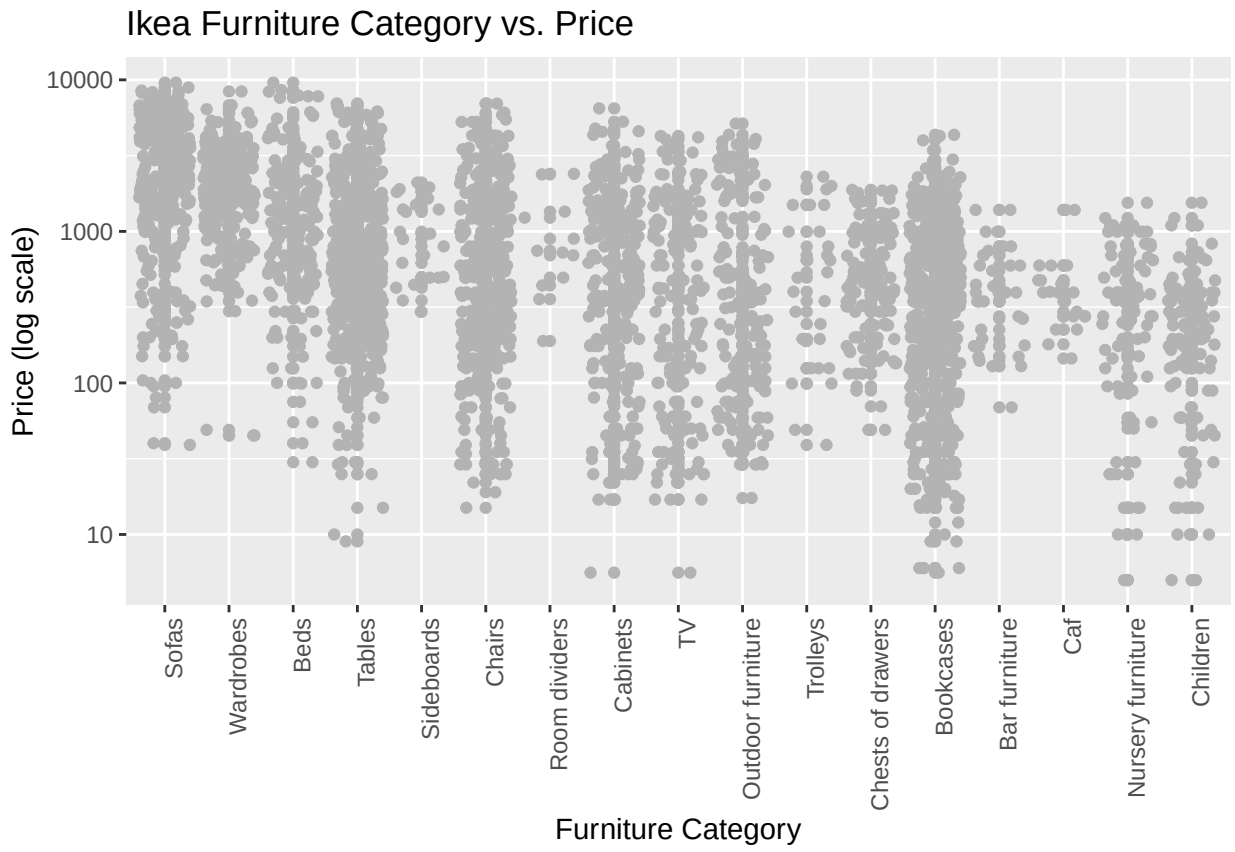
```
ggplot(ikea) +
  geom_point(aes(x = category, y = price), col = 'gray70') +
  scale_y_log10() +
  geom_jitter(aes(x = category, y = price), col = 'gray70') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    x = "Furniture Category",
    y = "Price (log scale)",
    title = "Ikea Furniture Category vs. Price"
  )
```



Part B

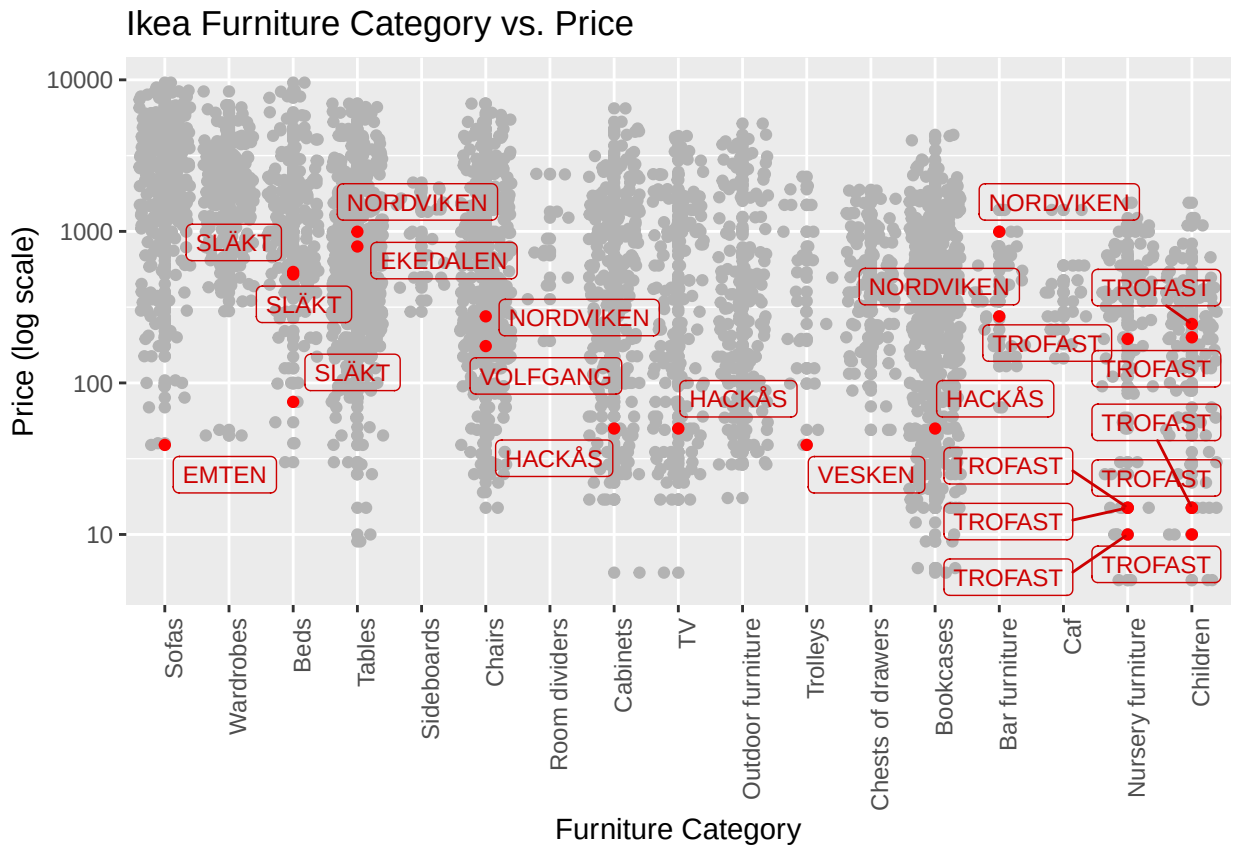
```
ikea$category = with(ikea, reorder(category, -price, mean))

ggplot(ikea) +
  geom_point(aes(x = category, y = price), col = 'gray70') +
  scale_y_log10() +
  geom_jitter(aes(x = category, y = price), col = 'gray70') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    x = "Furniture Category",
    y = "Price (log scale)",
    title = "Ikea Furniture Category vs. Price"
  )
```



Part C

```
highlight_df <- ikea %>%
            filter(sellable_online == F)
ggplot(ikea) +
  geom_point(aes(x = category, y = price), col = 'gray70') +
  scale_y_log10() +
  geom_jitter(aes(x = category, y = price), col = 'gray70') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_point(data = highlight_df, aes(x = category, y = price), color = 'red') +
  geom_label_repel(data = highlight_df, aes(x = category, y = price), label = highlight_df$name, col ='
  labs(
    x = "Furniture Category",
    y = "Price (log scale)",
    title = "Ikea Furniture Category vs. Price"
  )
```



(2) Penguins

The data below measures properties of various antarctic penguins.

```
penguins <- read_csv("https://uwmadison.box.com/shared/static/ijh7iipc9ect1jf0z8qa2n3j7dgem1gh.csv")
```

Using either vega-lite or ggplot2, create a single plot that makes it easy to answer both of these questions, i) How is bill length related to bill depth within and across species? ii) On which islands are which species found?

(3) 2012 London Olympics

This exercise is similar to the Ikea furniture one, except that it will be interactive. The data at this link describes all participants in the London 2012 Olympics. From an observable notebook, the following code can be used to derive a new variable with a jittered Age variable, which will be useful in part (a).

```
import { vl } from "@vega/vega-lite-api"
import { aq, op } from "@uwdata/arquero"
data_raw = aq.fromCSV(await FileAttachment("All London 2012 athletes - ALL ATHLETES.csv").text())
data = data_raw.derive({Age_: d => d.Age + 0.25 * Math.random() })
```

a. Create a layered display that shows (i) the ages of athletes across sports and (ii) the average age within each sport. Use different marks for participants and for averages. To avoid overplotting, use the jittered `Age_` variable defined in the code block above.

b. Sort the sports from lowest to highest average age. Add a tooltip so that hovering over an athlete shows their name. Your results should look something like the display in Figure **??**.

```
#\begin{figure}
 # \centering
  #\includegraphics[width=0.8\textwidth]{/Users/kris/Desktop/olympics.png}
  #\caption{Example vega-lite result for Problem (3). Hovering over a tick mark
  #shows the name of the athlete.}
  #\label{fig:3}
#\end{figure}
```

(4) Traffic

In lecture, we looked at the `geom_density_ridges` function. In this exercise, we will instead use `geom_ridgeline`, which is useful whenever the heights of the ridges have been computed in advance. We will use the traffic data read in below.

```
traffic <- read_csv("https://uwmadison.box.com/shared/static/x0mp3rhhic78vufsxtgrwencchmghbdf.csv")
```

Each row is a timepoint of traffic within a city in Germany. Using `geom_ridges`, make a plot of traffic over time, within each of the cities. An example result is shown below.

(5) Language Learning

This problem will look at a simplified version of the data from the study *A critical period for second language acquisition: Evidence from 2/3 million English speakers*, which measured the effect of the the age of initial language learning on performance in grammar quizzes. We have downloaded the raw data from the supplementary material and reduced it down to the average and standard deviations of test scores within (initial learning age) $\times$ (current age-group) combinations. We have kept a column `n` showing how many participants were used to compute the associated statistics. The resulting data are available here.

a) Using the `.derive()` command in arquero, create two new fields, `low` and `high`, giving confidence intervals for the means in each row. That is, derive new variables according to $\hat{x} \pm 2 * \frac{1}{\sqrt{n}}\hat{\sigma}$.
b) Create a `markArea`-based ribbon plot showing confidence intervals for average test scores as a function of starting age. Include a line for the average score within that combination. An example result is shown in Figure 5

```
#\begin{figure}
 # \centering
  #\includegraphics[width=0.6\textwidth]{/Users/kris/Desktop/languages.png}
  #\caption{Example result for problem (5).}
  #\label{fig:5}
#\end{figure}
```

(6) Deconstruction

Take a static screenshot from any of the visualizations in this article, and deconstruct its associated visual encodings.

a) What do you think was the underlying data behind the current view? What where the rows, and what were the columns?
b) What were the data types of each of the columns?
c) What encodings were used? How are properties of marks on the page derived from the underlying abstract data?
d) Is multi-view composition being used? If so, how?