479 HW 1

Caitlin Bolz

2/3/2021

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

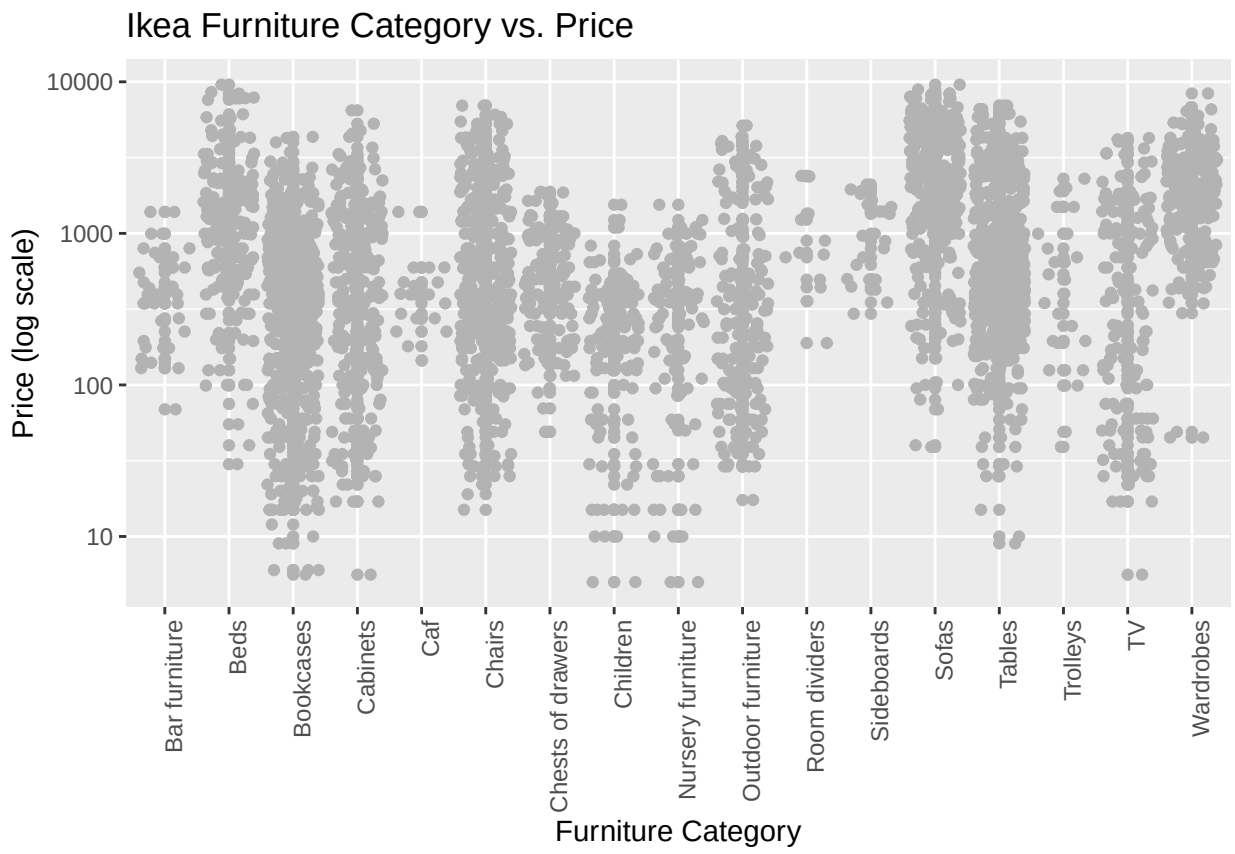
Problems

(1) Ikea Furniture

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

Part A

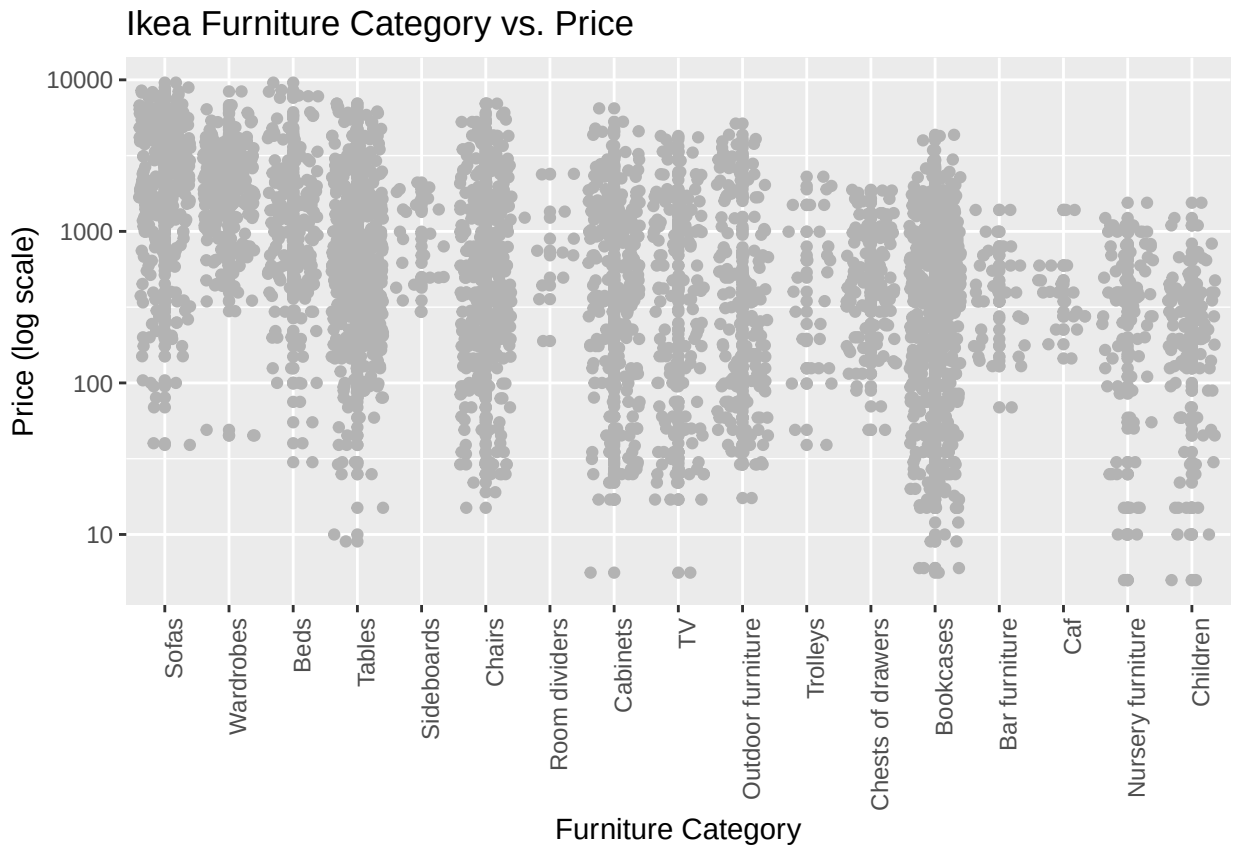
```
ggplot(ikea) +
  geom_point(aes(x = category, y = price), col = 'gray70') +
  scale_y_log10() +
  geom_jitter(aes(x = category, y = price), col = 'gray70') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    x = "Furniture Category",
    y = "Price (log scale)",
    title =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



Part B

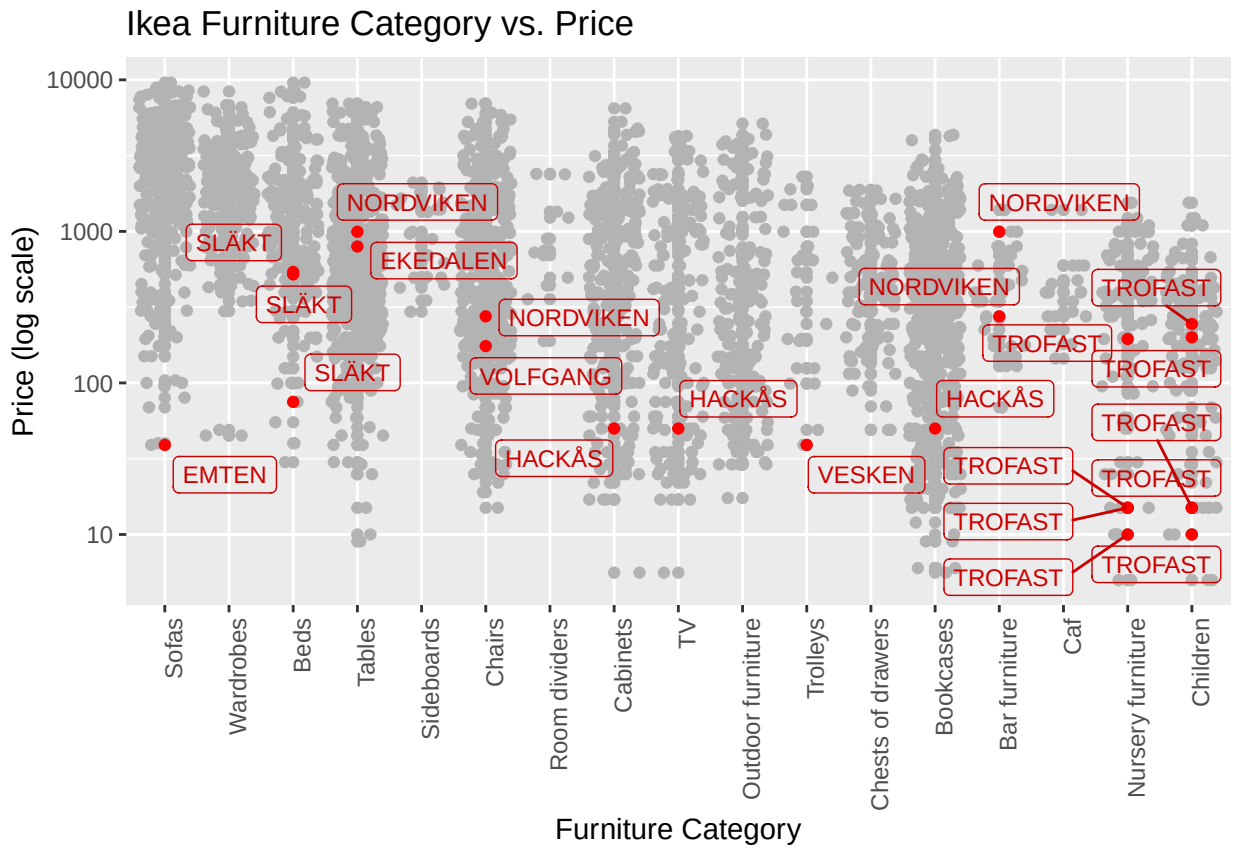
```
ikea$category = with(ikea, reorder(category, -price, mean))

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 ='red3', size = 3, force = 2,
                   max.overlaps = 20, fill = NA)+
  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")
```

ggplot(penguins, aes(bill_length_mm, y = species, fill = island)) + geom_density_ridges(alpha = .7) + scale_fill_brewer(palette = "Set2")

ggplot(gapminder_subset, aes(dollars_per_day, fill = group, y = ..count..)) + geom_density(alpha = 0.7, bw = .12, position = "stack") + scale_fill_brewer(palette = "Set2") + scale_x_log10(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0, 0.15, 0)) + facet_grid(year ~ .)

ggplot(penguins) + geom_point(aes(x = bill_length_mm, y = bill_depth_mm, col = species, group = island), alpha = 0.7, size = 0.9) + scale_color_brewer(palette = "Set2")

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.

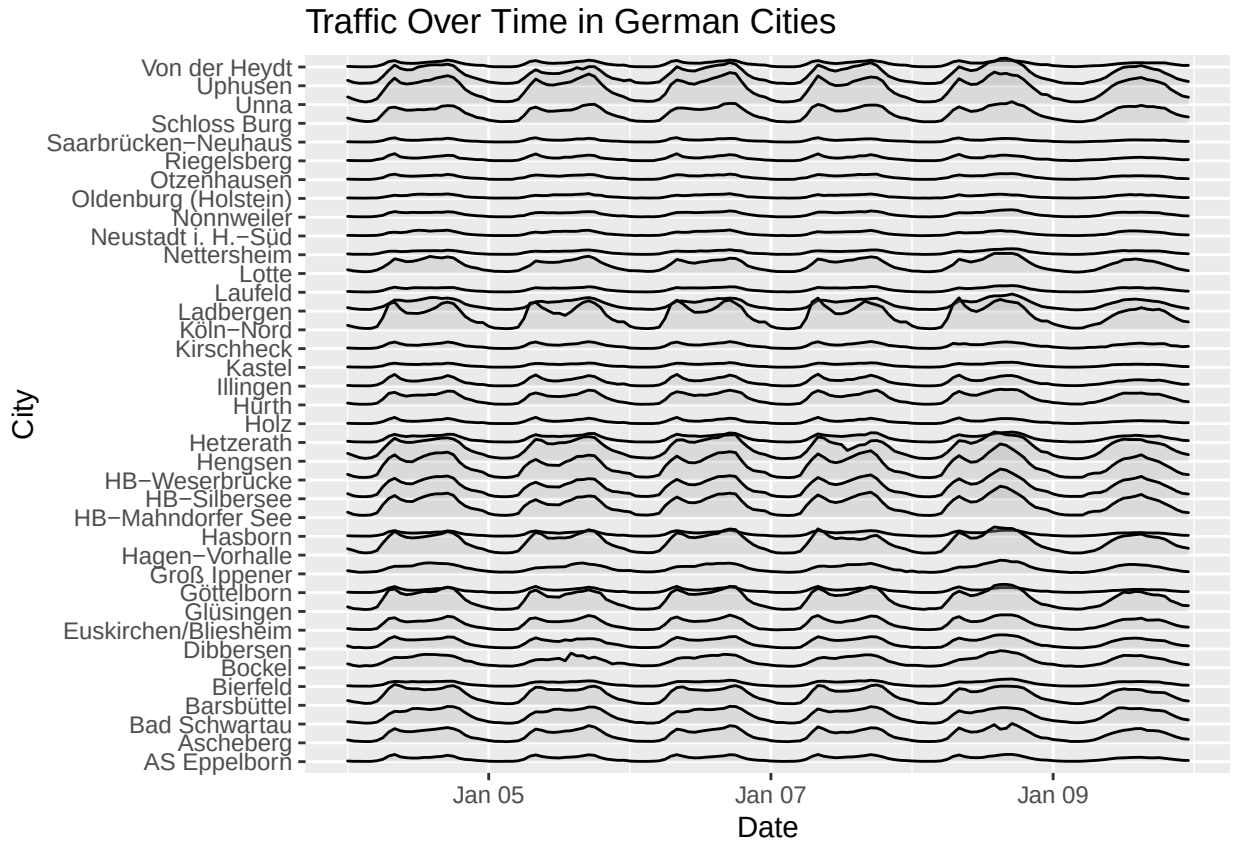
b. Sort the sports from lowest to highest average age. Add a tooltip so that hovering over an athlete shows their name. Your results should look something like the display in Figure **??**.

```
#\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

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

ggplot(traffic, aes(x = date, height = value, y = name)) +
  geom_ridgeline(scale= .4, alpha = .3)+
  labs(
    x = "Date",
    y = "City",
    title = "Traffic Over Time in German Cities"
  )
```



(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) × (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}
```

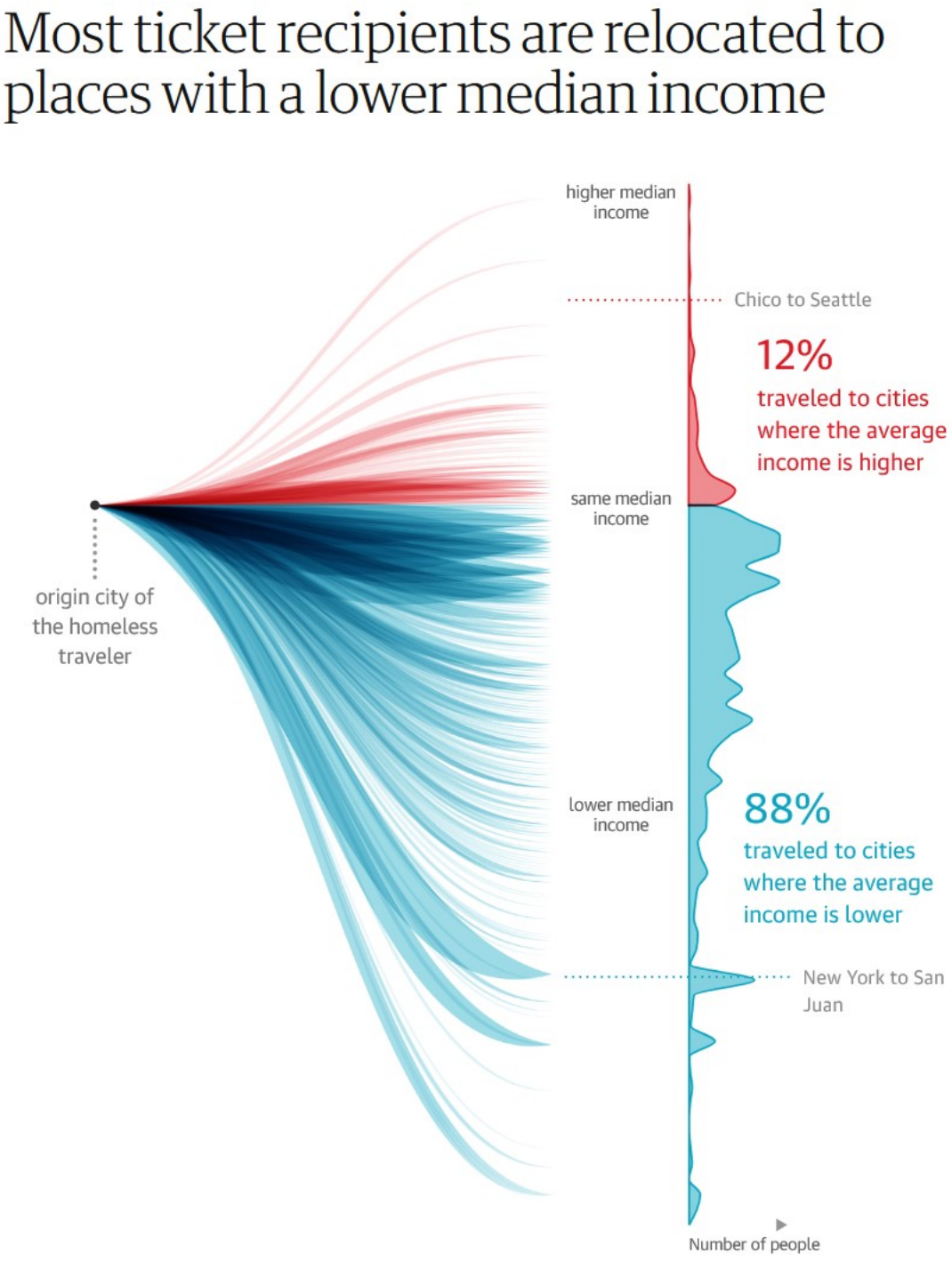


Figure 1: Alt text

(6) Deconstruction

Part A

Below are what I think the example columns are. There would be a row associated with each name/ID number 1. Names or ID numbers associated with each ticket recipient 2. Origin city of homeless traveler 3. Final Destination City 4. Average income of origin city 5. Average income of final city 6. Difference in income in two cities

Part B

Listed below are the data types for the columns described in Part A 1. Character or integer 2. Character 3. Character 4. Numeric 5. Numeric 6. Numeric

Part C

c) What encodings were used? How are properties of marks on the page derived from the underlying abstract data?

All of the origin cities start at the same point, giving the illusion that everyone starts in the same city. But really the origin city income is being compared to increase or decrease in the final destination city income.

The difference in median income between the cites is encoded in in the lines we see from the origin point to the left. The count for each difference in median income is displayed through a density plot on the far right.

rows are marks columns are properties of the marks

Nominal – cities, name/id quantitative – prices

Part D

Yes. The visualization shows if people went to cities that had a lower or higher median income. In addition, it also shows the density of how many people made that specific increase or decrease in median income.