Lab 3

Emmanuel Maduneme

2023-02-13

Getting Started

You can download the `transit_cost.csv` data from the website.

```
require(tidyverse)
require(lubridate)
require(ungeviz)
require(ggtext)
require(ggrepel)
require(ggforce)
require(pacman)
p_load(here, rio, skmir, ggthemes)

transit_cost <- read_csv(here("data", "transit_cost.csv"))
```

Question 1

Suppose that you want to demonstrate the relationship between Average Length and Average Cost for the transit systems across all countries in the dataset. Reproduce the plot on the next page by following the procedures:

1. Compute the average length and average cost of transit systems by country and city

2. Create a basic scatter plot by placing **Average Length** on the x-axis and **Average Cost** on the y-axis.

3. In the scatter plot,make the size of the data points represent the number of transit systems in that particular city (Hint: use `aes(size=)` within the `geom_point()` function).

4. Customize the legend so it shows 5, 10, and 20 as break points for the size of data points (hint: add the feature to the plot by using `scale_size_binned()`)

5. Make sure all data points are grayish except the cities from India. Make the color for the data points from these 9 cities different than the rest.

6. Adjust the scale of the x-axis and y-axis using the `scale_y_log10()` and `scale_x_log10()` functions so they are on the logarithmic scale.

7. Add the names of the cities in India using the `geom_text_repel()` function.

8. Adjust the theme settings.

```r
p_load(gghighlight)


d <- transit_cost %>%
  group_by(city, country) %>%
  summarize(avg_length = mean(length),
            avg_cost = mean(as.numeric(real_cost), na.rm = T),
            count_city = n())


d %>%
ggplot(aes(avg_length,
           avg_cost)) +
  geom_point(aes(size = count_city), col = "Purple")+
  gghighlight(country == "IN") +
  scale_x_log10(limits = c(1, 100)) +
  scale_y_log10(limits = c(1000, 100000), labels= scales::dollar) +
  scale_size_binned(name = "Number of Transit", breaks = c(5, 10, 20)) +
  geom_text_repel(aes(label = city)) +
  scale_fill_manual(values = "Purple") +
  theme_bw() +
  labs(title = "Longer transit systems tend to cost more",
       subtitle = "India has among the most transit systems in the world",
       x = "Average length",
       y = "Average Cost") +
  theme(legend.position = "bottom",
        legend.direction = "horizontal",
        legend.title = element_text(size = 10, face = "bold"),
        legend.key.size = unit(.5, 'cm'),
        legend.key.height = unit(.5, "cm"))
```

Longer transit systems tend to cost more
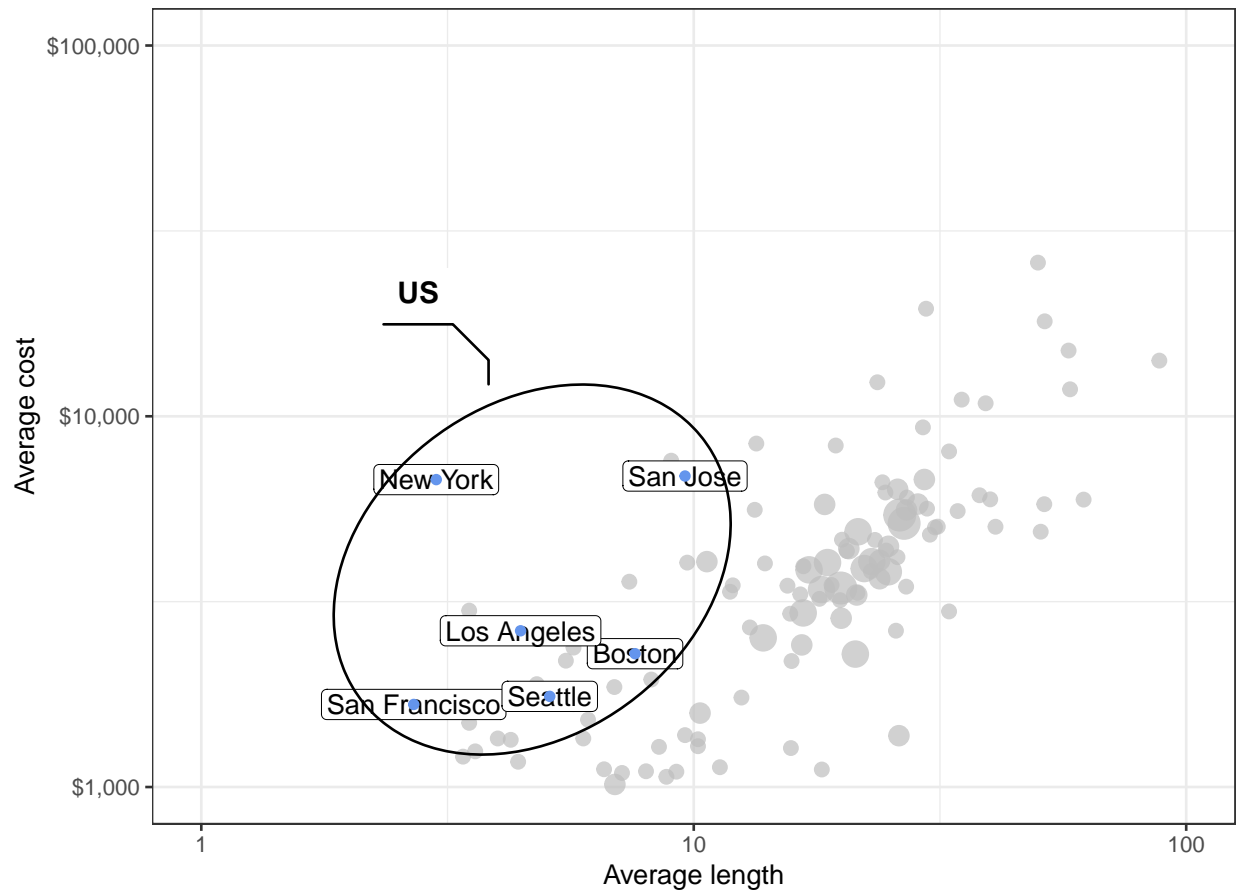India has among the most transit systems in the world

Question 2

Using basically the same data, reproduce the following plot on the next page.

1. Compute the average length and average cost of transit systems by country and city.

2. Create a basic scatter plot by placing **Average Length** on the x-axis and **Average Cost** on the y-axis.

3. In the scatter plot,make the size of the data points represent the number of transit systems in that particular city (Hint: use `aes(size=)` within the `geom_point()` function).

4. Customize the legend so it shows 5, 10, and 20 as break points for the size of data points (hint: add the feature to the plot by using `scale_size_binned()`)

5. Make sure all data points are grayish except the cities from US. Make the color for the data points from the US cities different than the rest.

6. Adjust the scale of the x-axis and y-axis using the `scale_y_log10()` and `scale_x_log10()` functions so they are on the logarithmic scale.

7. Using the `geom_mark_ellipse()` function from the `ggforce` package, circle the data points for the US cities.

8. Add the names of the US cities using the `geom_label_repel()` function.

9. Adjust the theme settings.

```
d %>%
  ggplot(aes(avg_length, avg_cost)) +
  geom_point(aes(size = count_city), col = "grey70") +
  scale_size_binned(breaks = c(5, 10, 20), name = "Number of transit systems") +
  gghighlight(city %in% c("New York", "Los Angeles",
                          "San Francisco", "Seattle", "Boston", "San Jose")) +
   scale_x_log10(limits = c(1, 100)) +
  scale_y_log10(limits = c(1000, 100000), labels= scales::dollar) +
  geom_label(aes(label = city),
                  label.padding = unit(0.15, "lines"),
                  label.size = 0.05,
                  stat = "unique") +
  geom_point(aes(label = city),
             color = "cornflowerblue") +
  geom_mark_ellipse(aes(label = country)) +
  labs(title = "Longer transit system tend to cost more",
       subtitle = "United States has most expensive transit systems (average cost per average length)",
       x = "Average length",
       y = "Average cost") +
  theme_bw() +
  theme(
       legend.position = "bottom",
       legend.direction = "horizontal",
       legend.title = element_text(size = 15, face = "bold"),
       legend.key.size = unit(1, 'cm'),
       legend.key.height = unit(1, "cm"))
```

Longer transit system tend to cost more
United States has most expensive transit systems (average cost per average length)

US

New York   San Jose

Los Angeles
Boston
San Francisco Seattle

Average cost

$100,000

$10,000

$1,000

1   10   100

Average length

Number of transit systems

5   10   20