Lab 3

Seulbi Lee

2023-02-11

Getting Started

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

```r
require(tidyverse)
require(lubridate)
require(ungeviz)
require(ggtext)
require(ggrepel)
require(ggforce)
library(dplyr)

df <- read_csv('./transit_cost.csv') %>% drop_na()
```

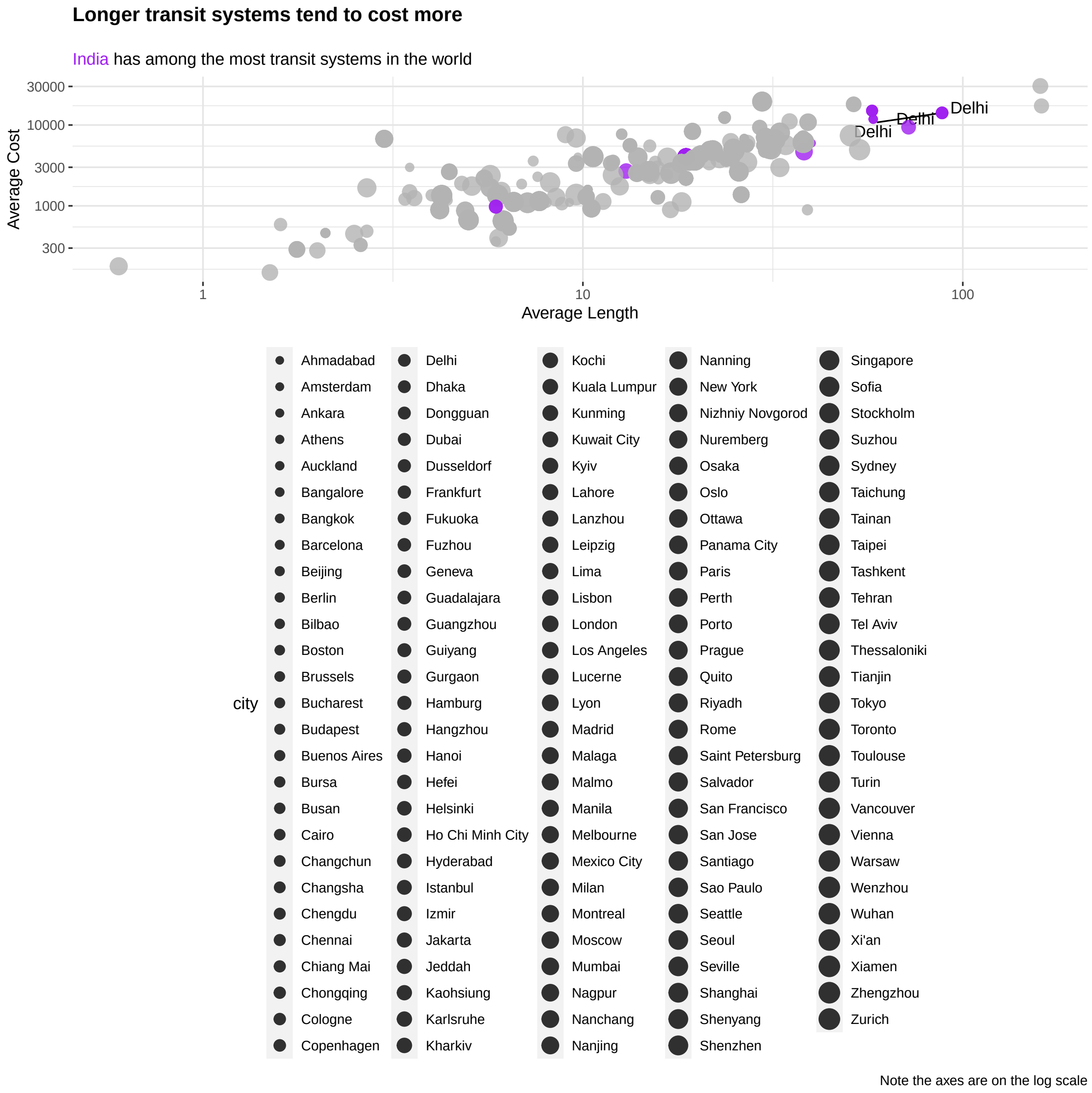
Question 1

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

df$length <- as.numeric(df$length)

length_city <- df |>
    group_by(country, city) |>
    dplyr::summarise(city_avg_length = mean(length))

length_country <- df |>
    group_by(country) |>
    dplyr::summarise(country_avg_length = mean(length))

df1 <- left_join(length_city, length_country)

df$real_cost <- as.numeric(df$real_cost)

cost_city <- df |>
    dplyr::group_by(country, city) |>
    dplyr::summarise(city_avg_cost = mean(real_cost))

cost_country <- df |>
    dplyr::group_by(country) |>
    dplyr::summarise(country_avg_cost = mean(real_cost))

df2 <- left_join(cost_city, cost_country)

library(plyr)
df <- join_all(list(df, df1,df2), type='left')
```

```r
# Ver.1
# Looks okay except for the legend, so I try another.

library(ggtext)
lab <- "<span style='color:#A020F0;'>India</span>has among the most transit systems in the world"


p1 <-
  ggplot(df, aes(x=city_avg_length, y=city_avg_cost,
                 label=ifelse(country == "IN", city, ""))) +
  geom_text_repel(stat="identity")+
  geom_point(aes(size=city), alpha=.8,
             color=ifelse(df$country == "IN","purple", "grey70")) +
  scale_x_log10()+
  scale_y_log10()+
  labs(title="**Longer transit systems tend to cost more**<br>",
       subtitle="<span style='color:#A020F0;'>India</span> has among the most transit systems in the wo
       x="Average Length", y="Average Cost",
       caption="Note the axes are on the log scale") +
  theme(legend.position='bottom', panel.background=element_rect(fill="white"),
        panel.grid.major = element_line(colour = "grey90"), panel.grid.minor =element_line(colour = "gr
        plot.title=element_markdown(), plot.subtitle=element_markdown())

p1
```

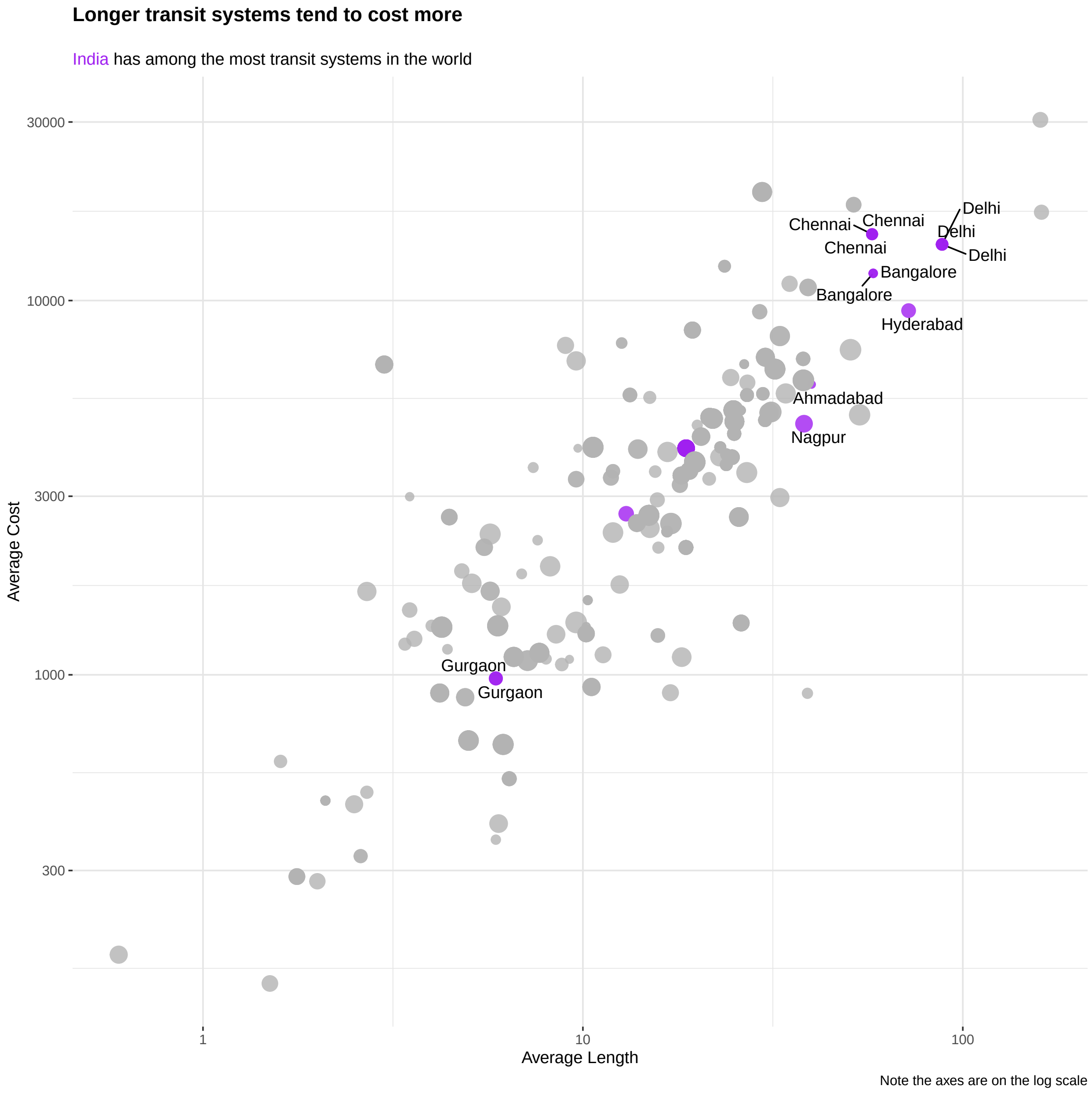


Longer transit systems tend to cost more

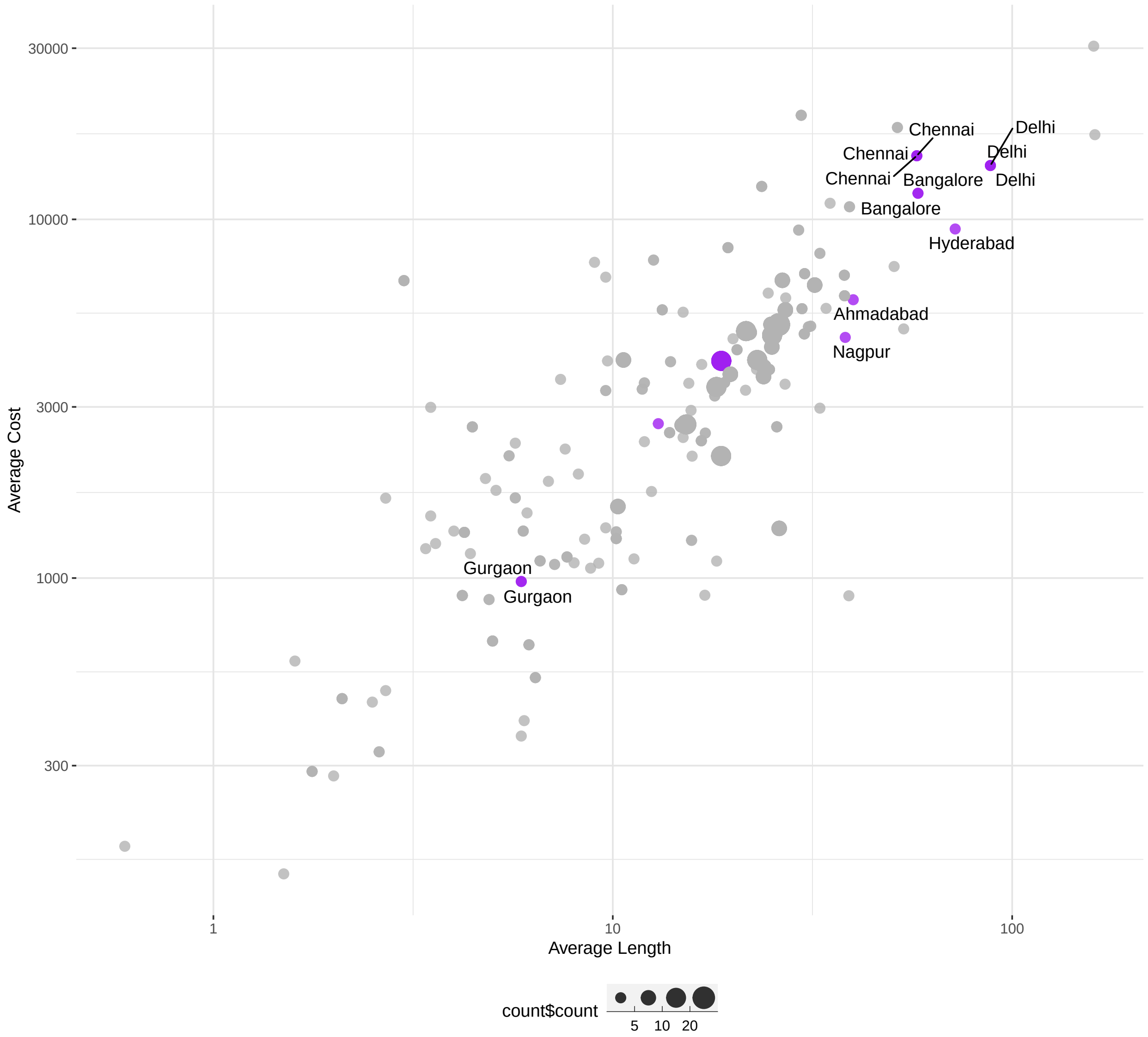
India has among the most transit systems in the world

Note the axes are on the log scale

```r
p1 + theme(legend.position = 'none')
```



```
# Ver. 2
# This time, I just made another column for the number of transit per city. It looks okay, but I don't

df$count <- df %>%
  dplyr::group_by(city) %>%
  dplyr::mutate(count = n())

options(scipen=999)

p2 <-
  ggplot(df, aes(x=city_avg_length, y=city_avg_cost,
               label=ifelse(country == "IN", city, ""))) +
  geom_point(aes(size=count$count), alpha=.8,
             color=ifelse(df$country == "IN","purple", "grey70")) +
  geom_text_repel()+
  scale_size_binned(breaks = c(5, 10, 20))+
  scale_x_log10()+
  scale_y_log10()+
  # scale_y_continuous()+
  labs(title="**Longer transit systems tend to cost more**<br>",
       subtitle="<span style='color:#A020F0;'>India</span> has among the most transit systems in the wo
       x="Average Length", y="Average Cost",
       caption="Note the axes are on the log scale") +
  theme(legend.position='bottom', panel.background=element_rect(fill="white"),
        panel.grid.major = element_line(colour = "grey90"), panel.grid.minor =element_line(colour = "gr
        plot.title=element_markdown(), plot.subtitle=element_markdown())


p2
```



Longer transit systems tend to cost more

India has among the most transit systems in the world

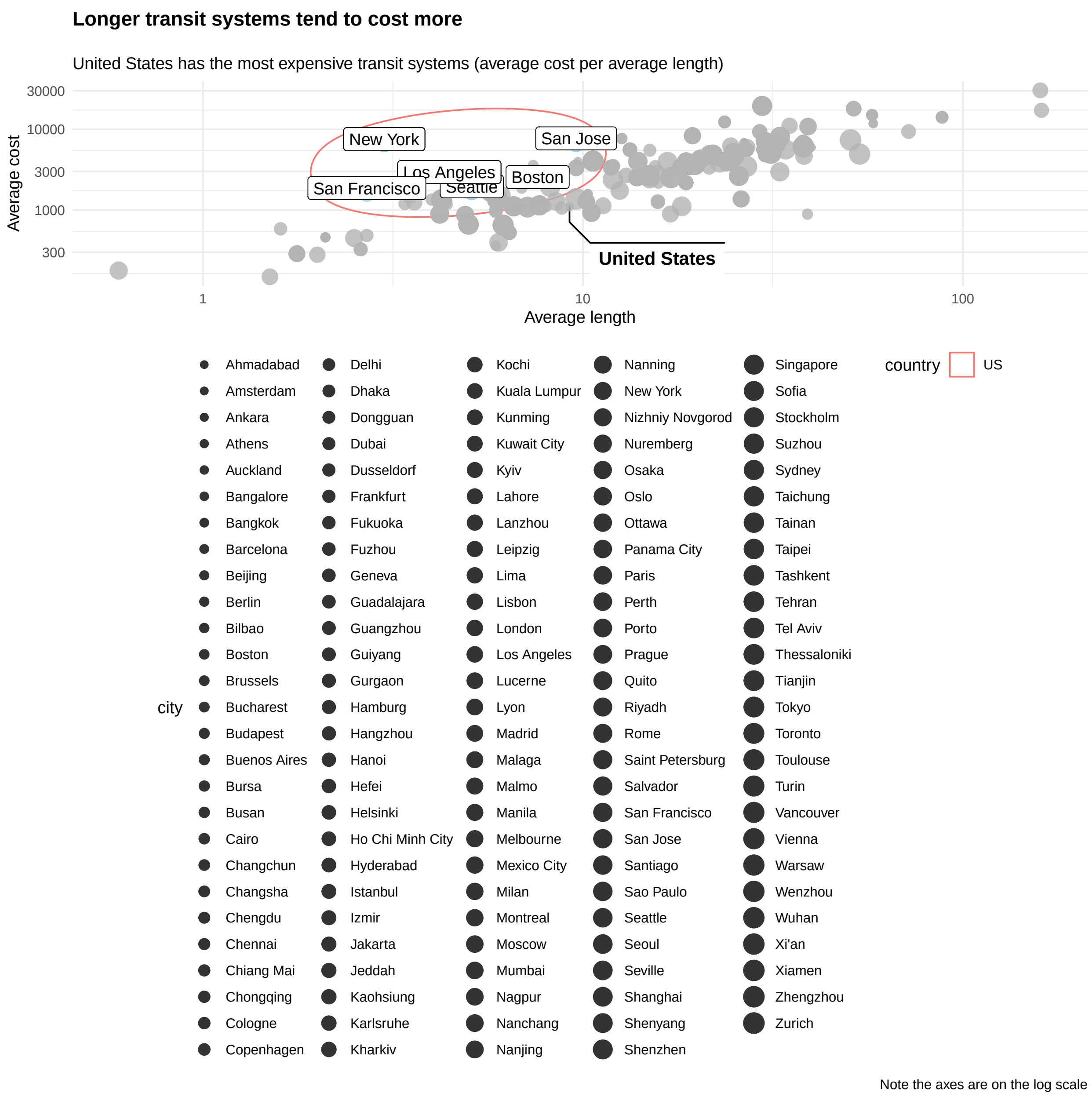
Note the axes are on the log scale

Question 2

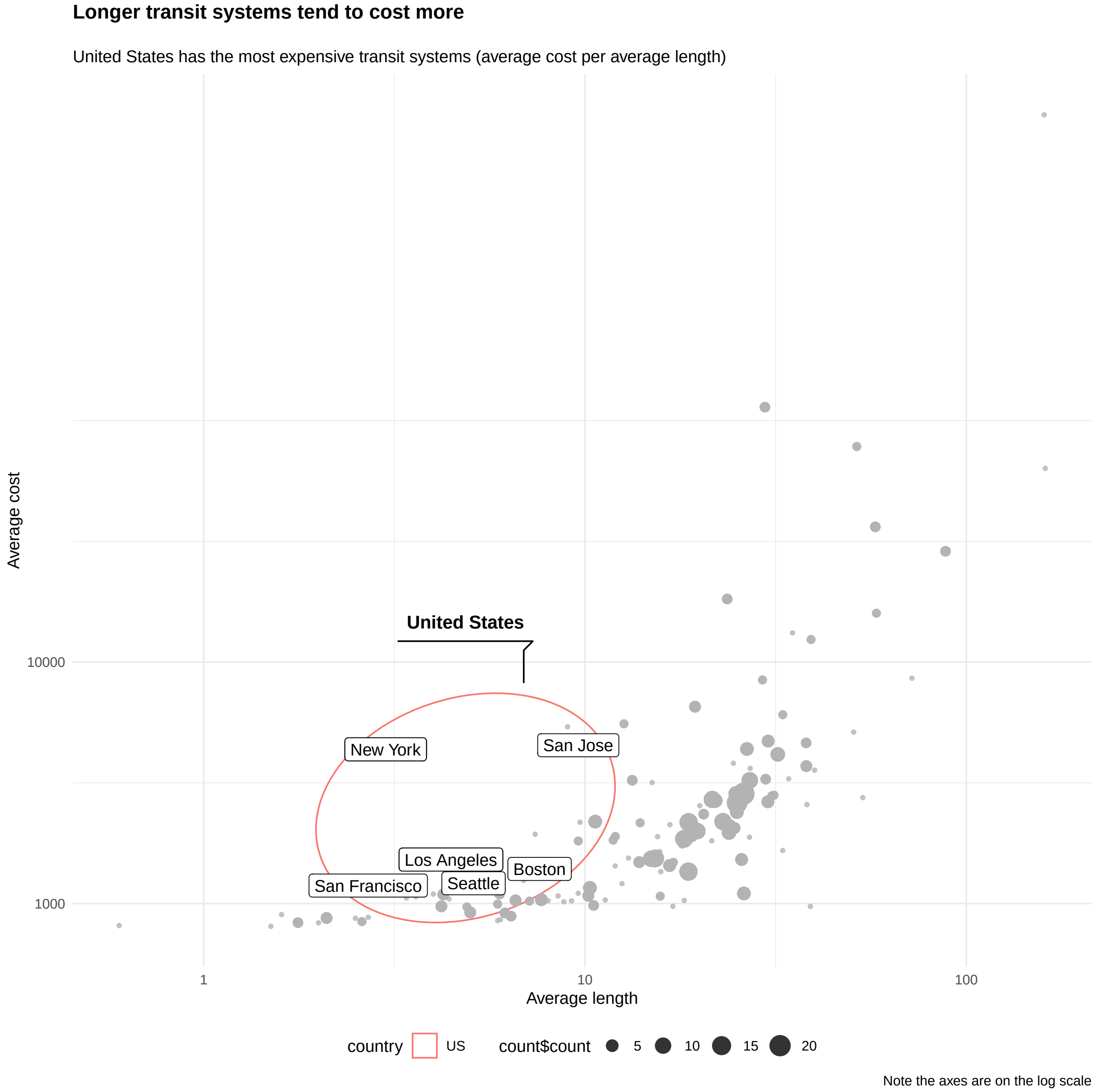
Using basically the same data, reproduce the following plot on the next page.

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.

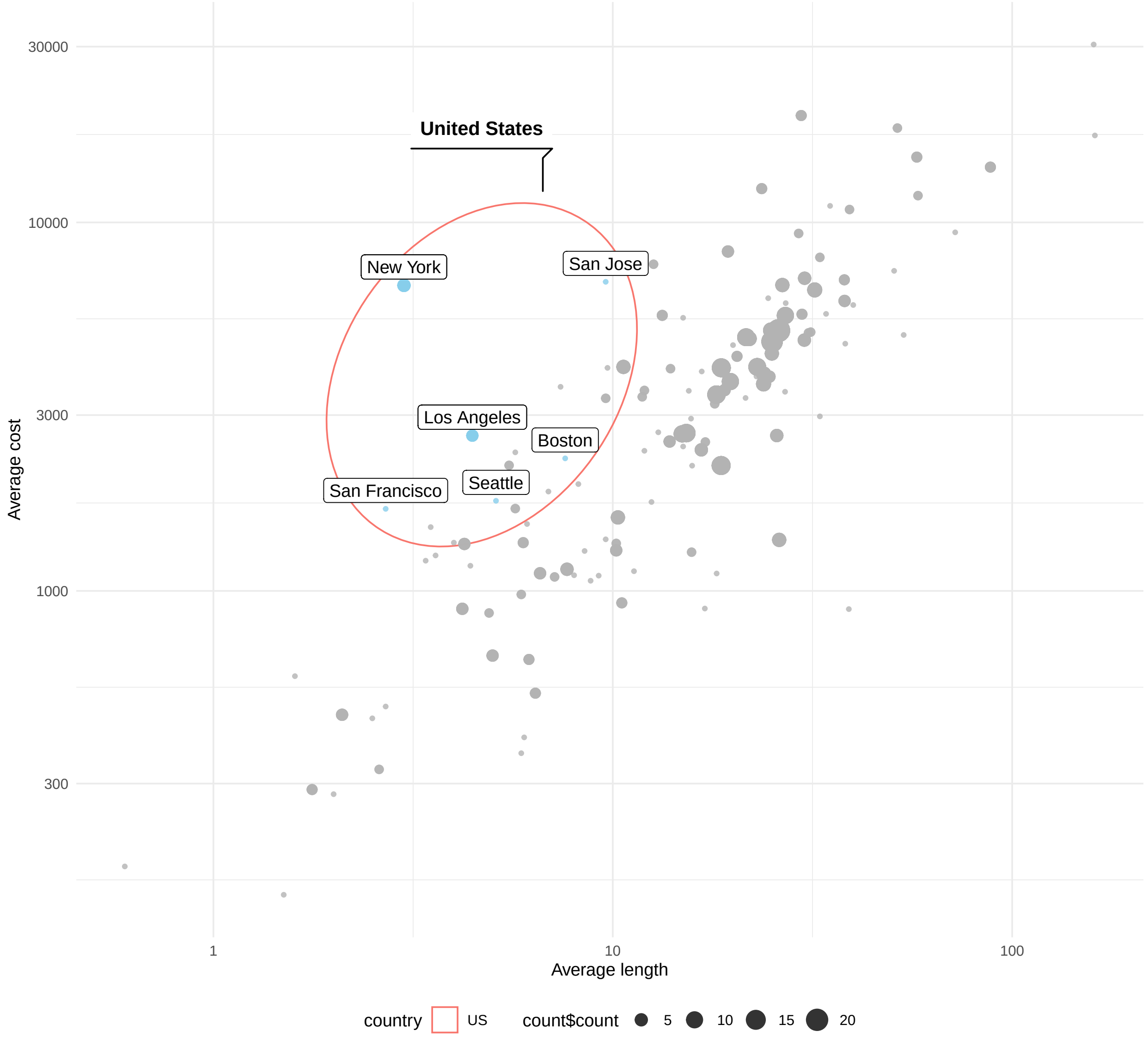
```r
# Ver.1
# Everything works but legend is not like what I want again

p3 <-
  ggplot(df, aes(x=city_avg_length, y=city_avg_cost,
                 label=ifelse(country == "US", city, ""))) +
  # geom_text_repel()+
  geom_mark_ellipse(aes(color=country, label="United States", filter = country == "US")) +
  geom_point(aes(size=city), alpha=.8,
             color=ifelse(df$country == "US","skyblue", "grey70")) +
  scale_x_log10()+
  scale_y_log10()+
  geom_label(data=df |> dplyr::filter(country == "US"),
             aes(label=city),
             check_overlap = TRUE, nudge_y = 0.05) +
  labs(title="**Longer transit systems tend to cost more**<br>",
       subtitle="United States has the most expensive transit systems (average cost per average length)
       x="Average length", y="Average cost", caption="Note the axes are on the log scale")+
  theme_minimal()+
  theme(plot.title=element_markdown(), legend.position = 'bottom')

p3
```



```
# Ver.2

p4 <-
  ggplot(df, aes(x=city_avg_length, y=city_avg_cost,
                 label=ifelse(country == "US", city, ""))) +
  # geom_text_repel()+
  geom_mark_ellipse(aes(color=country,label="United States", filter = country == "US")) +
  geom_point(aes(size=count$count), alpha=.8,
             color=ifelse(df$country == "US","skyblue", "grey70")) +
  scale_x_log10()+
  scale_y_log10()+
  geom_label(data=df |> dplyr::filter(country == "US"), aes(label=city), check_overlap = TRUE, nudge_y
  labs(title="**Longer transit systems tend to cost more**<br>",
       subtitle="United States has the most expensive transit systems (average cost per average length)
       x="Average length", y="Average cost", caption="Note the axes are on the log scale")+
  theme_minimal()+
  theme(plot.title=element_markdown(), legend.position = 'bottom')


# This would make y axis as we want, but so much space above
p4 + scale_y_continuous(breaks=c(1000,10000,100000))
```



```
# scale_y_log() doesn't seem compatible with scale_y_continuous
# that would help me setting the y axis break
p4 + scale_y_continuous(breaks=c(1000,10000,100000)) +
  scale_y_log10()
```



Longer transit systems tend to cost more

United States has the most expensive transit systems (average cost per average length)

Note the axes are on the log scale