Lab 3

Tian Walker

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(rio)
require(here)
require(janitor)

#transit_cost <- read_csv('./transit_cost.csv')
transit_cost <- import(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

```
t1 <- transit_cost %>%
group_by(country, city) %>%
summarize(av_length = mean(length, na.rm = T), av_cost = mean(cost, na.rm = T))

t1
```

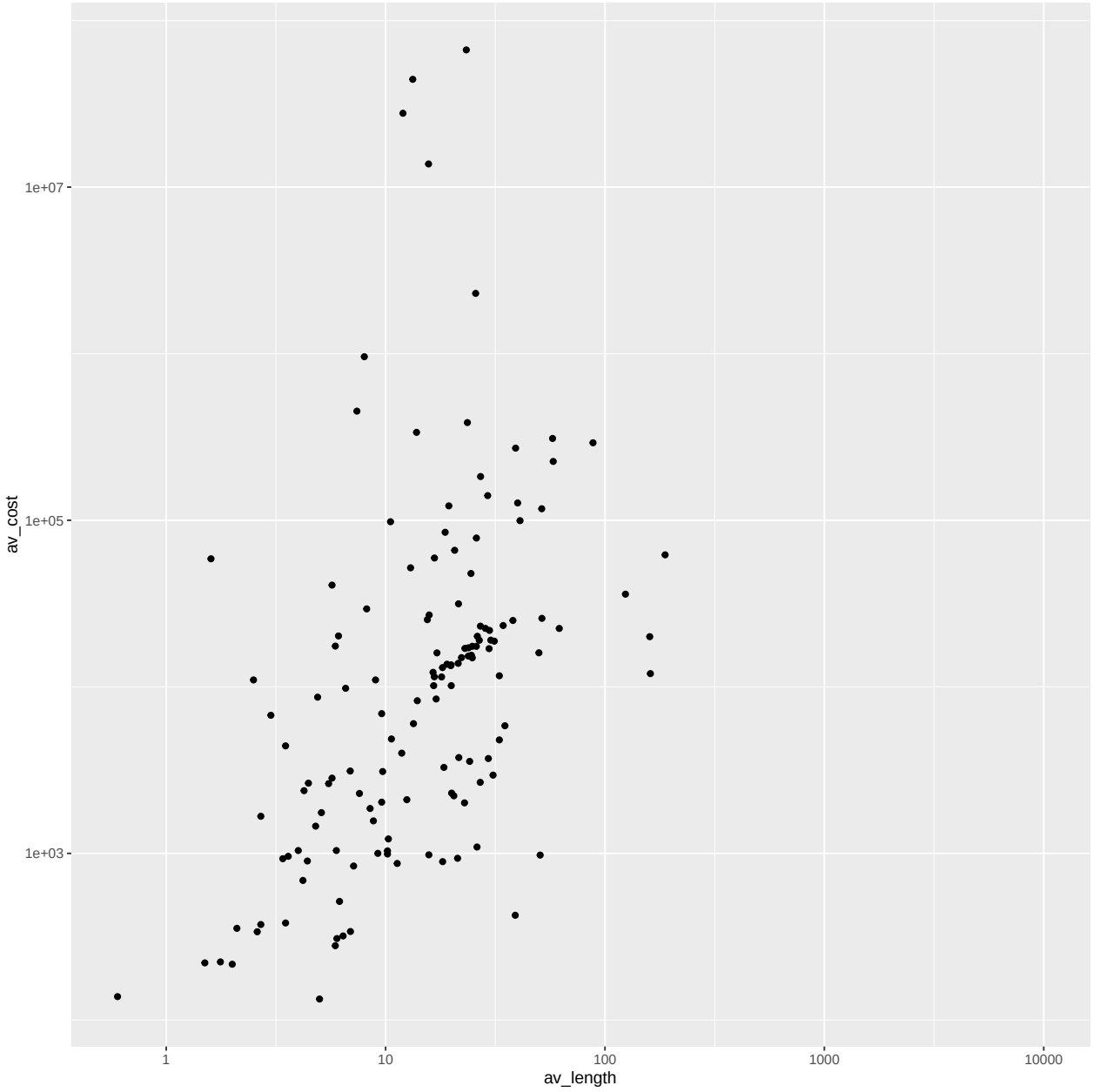
```
## # A tibble: 141 x 4
## # Groups:   country [57]
##    country city         av_length av_cost
##    <chr>   <chr>            <dbl>   <dbl>
##  1 AE      Dubai             24.2    3567.
##  2 AR      Buenos Aires      20      2300
##  3 AT      Vienna             5.97   1040
##  4 AU      Melbourne          9     11000
##  5 AU      Perth              8.5    1860
##  6 AU      Sydney            33     11650
##  7 BD      Dhaka             23.6  385997.
##  8 BE      Brussels           4.4     900
##  9 BG      Sofia              6.92    340.
## 10 BH      Bahrain           50.7     976.
```

```
## # ... with 131 more rows
```

```
#getting a closer look at the outliers (prior to log transformation)
newdata <- t1[order(t1$av_cost),]
```

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

```
d <- transit_cost %>%
group_by(city) %>%
summarize(av_length = mean(length, na.rm = T),
          av_cost = mean(cost, na.rm = T))



ggplot(d,aes(av_length, av_cost)) +
  geom_point(size = 1.5)+
  scale_x_log10() +
  scale_y_log10()
```

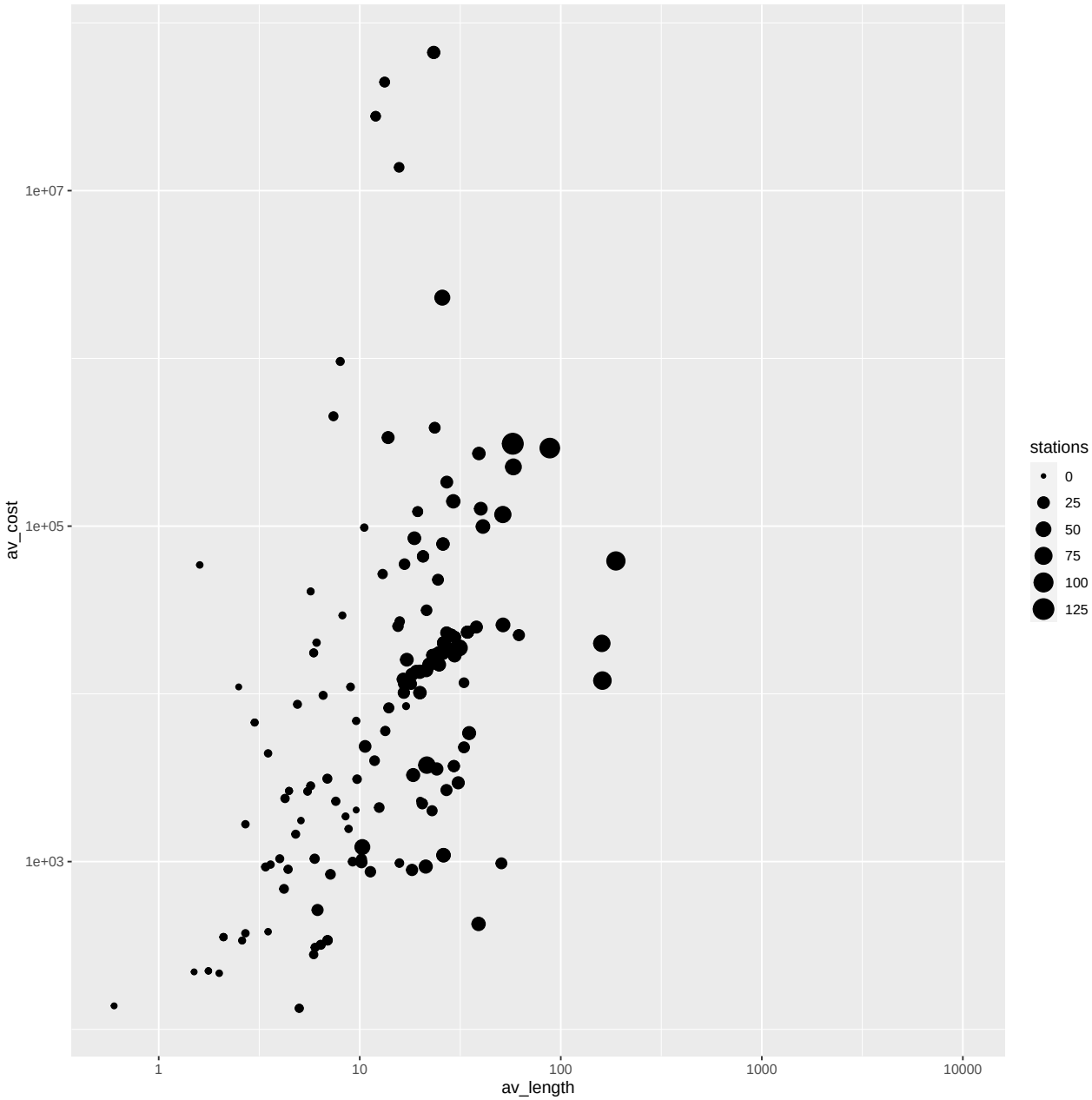



```
#use log
```

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).

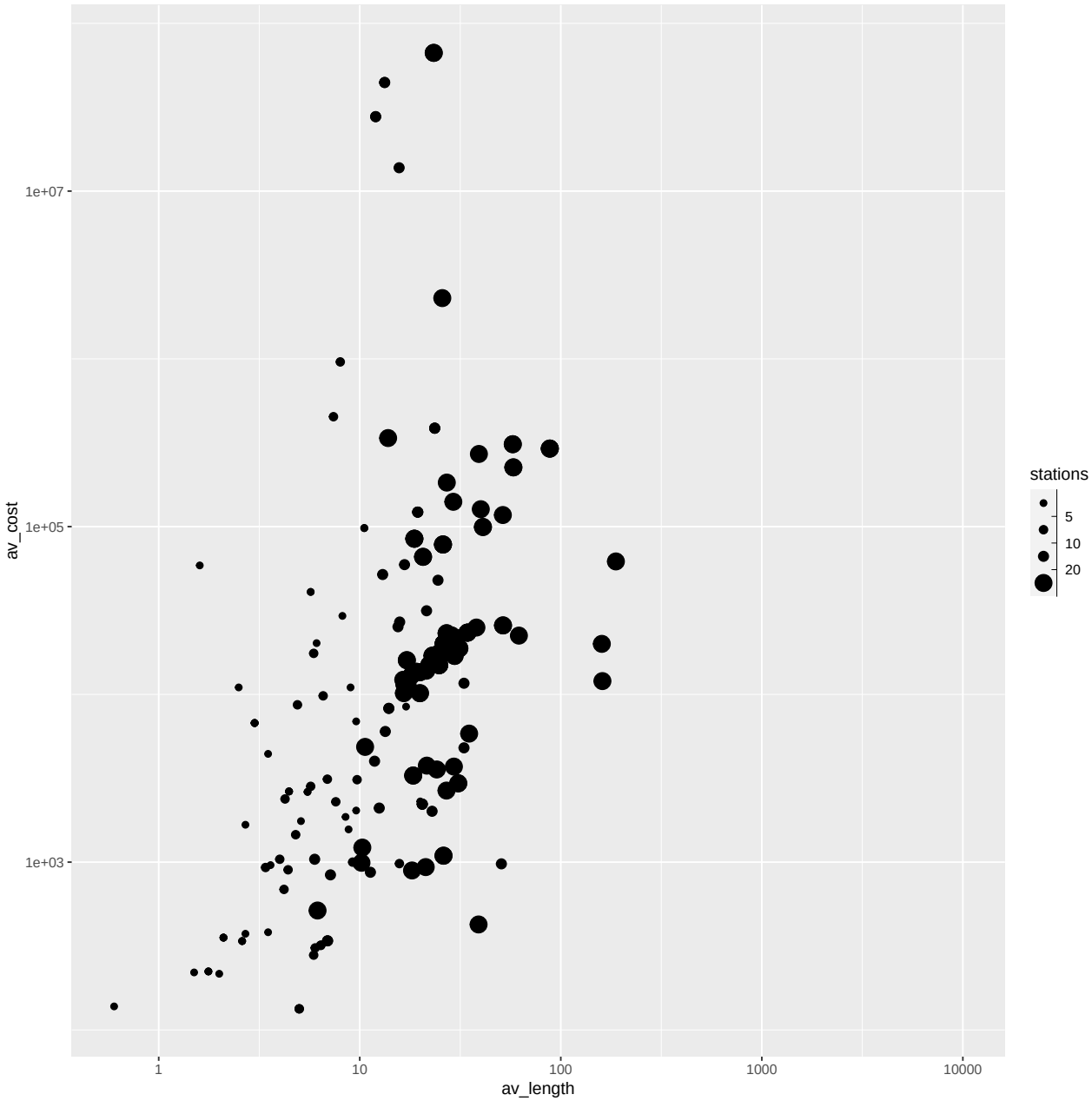
```
d2 <- transit_cost %>%
group_by(city) %>%
summarize(av_length = mean(length, na.rm = T),
          av_cost = mean(cost, na.rm = T), stations)



ggplot(d2,aes(av_length, av_cost)) +
  geom_point(aes(size = stations)) +
  scale_x_log10() +
```

```
  scale_y_log10()
```



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()`)

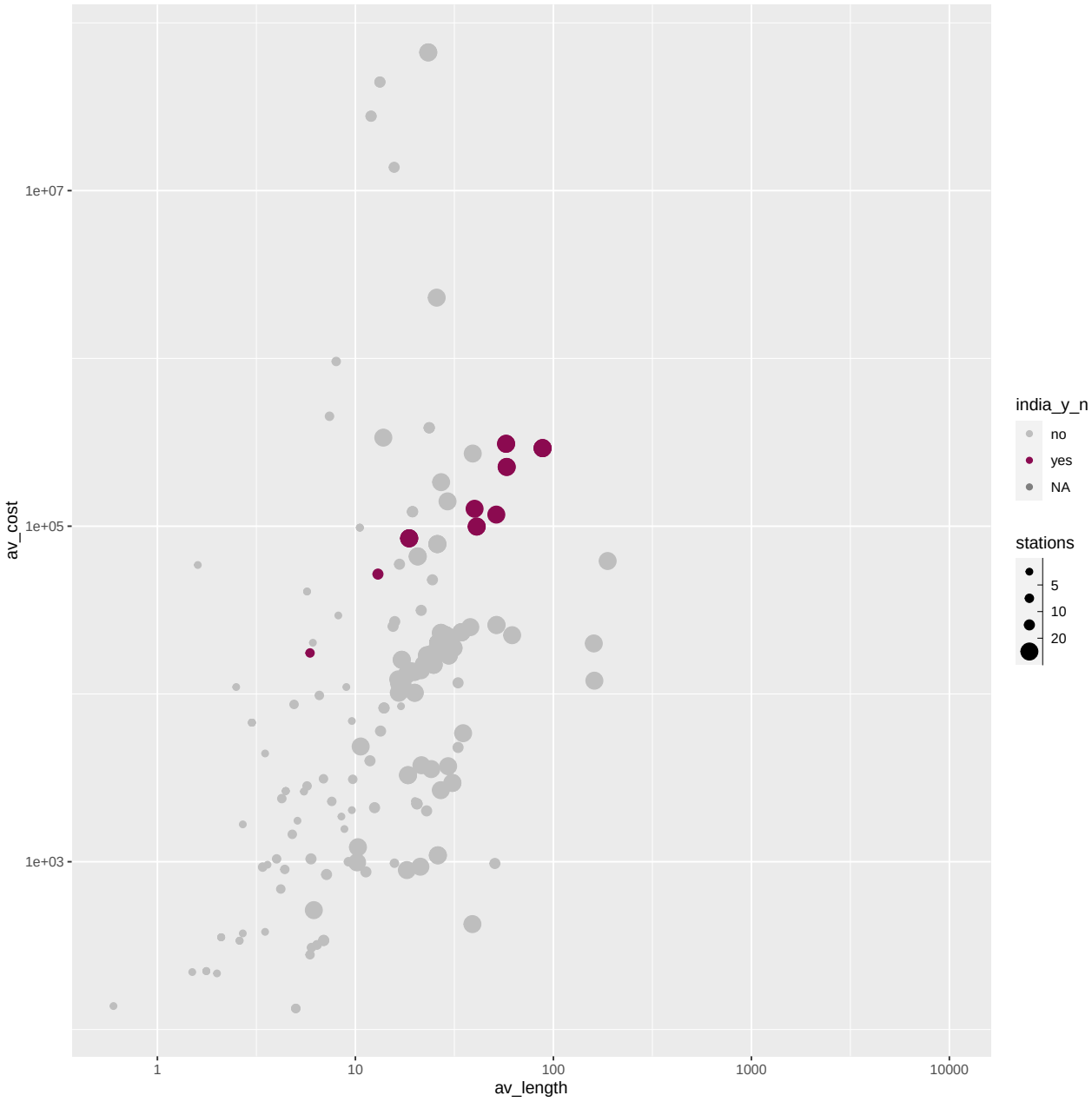
```
ggplot(d2,aes(av_length, av_cost)) +
  geom_point(aes(size = stations)) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20))
```



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.

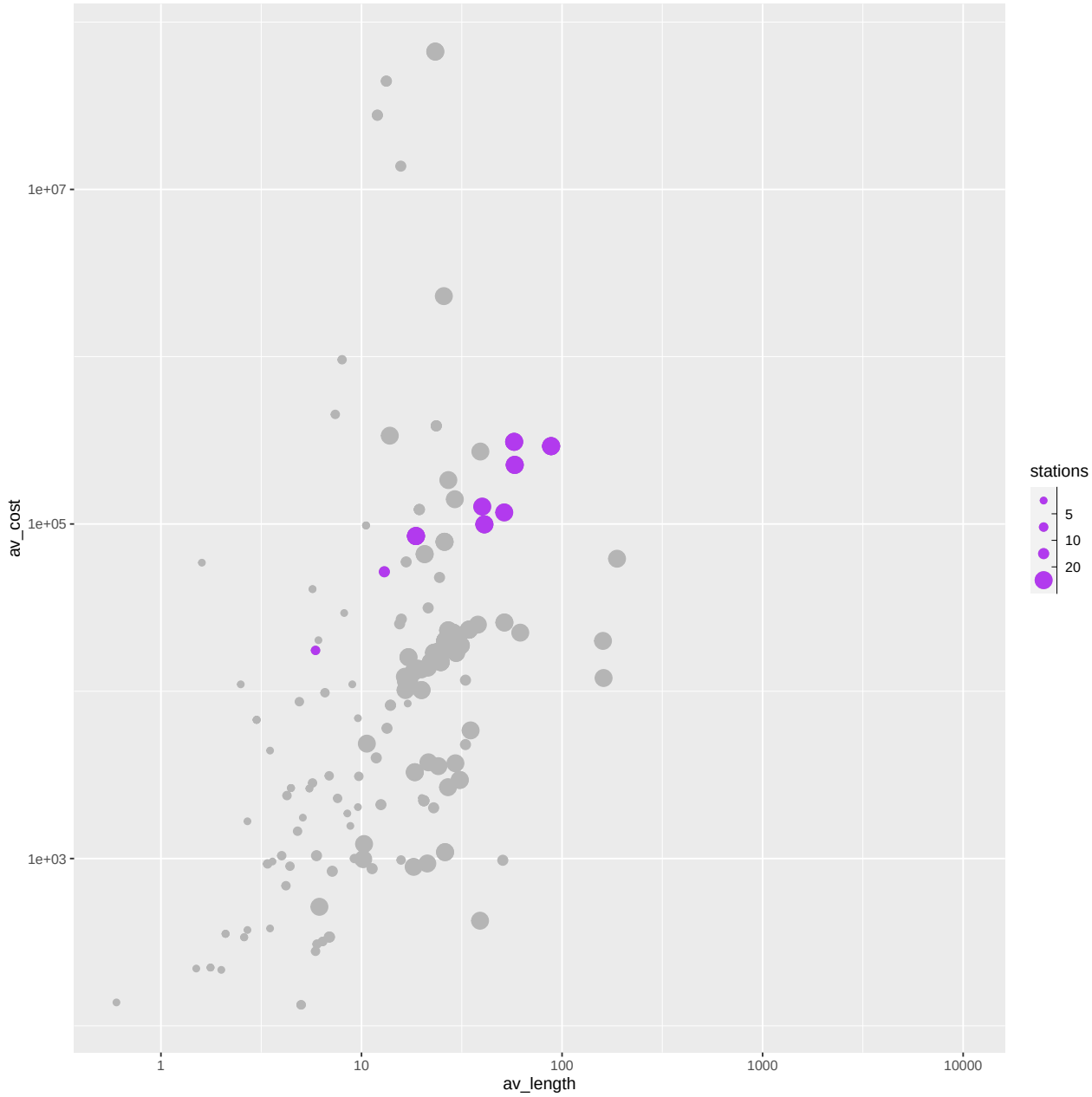
```
d3 <- transit_cost %>%
 mutate(india_y_n = ifelse(country == "IN", "yes", "no" ))


d3 <- d3 %>%
group_by(city) %>%
summarize(av_length = mean(length, na.rm = T),
          av_cost = mean(cost, na.rm = T), stations, india_y_n, country)

d3.0 <- d3 %>%
  group_by(stations) %>%
```

```
 filter(country  == "IN")


#option 1
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations, color = india_y_n)) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20)) +
  scale_color_manual (values = c("gray", "deeppink4"))
```



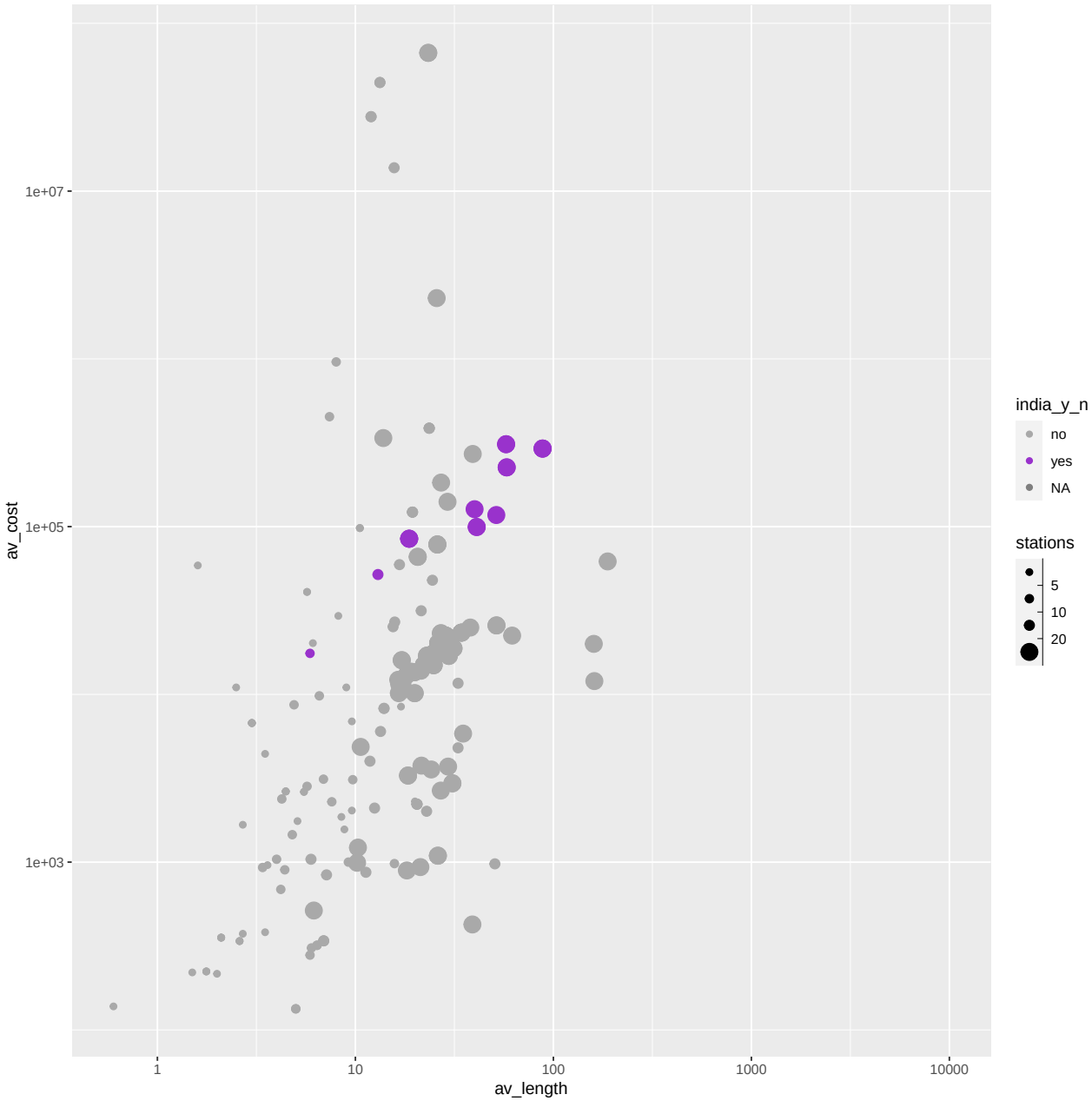

```
#option 2
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
```

```
  geom_point(data = d3.0, aes(size = stations), color = "darkorchid2") +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20))
```



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.

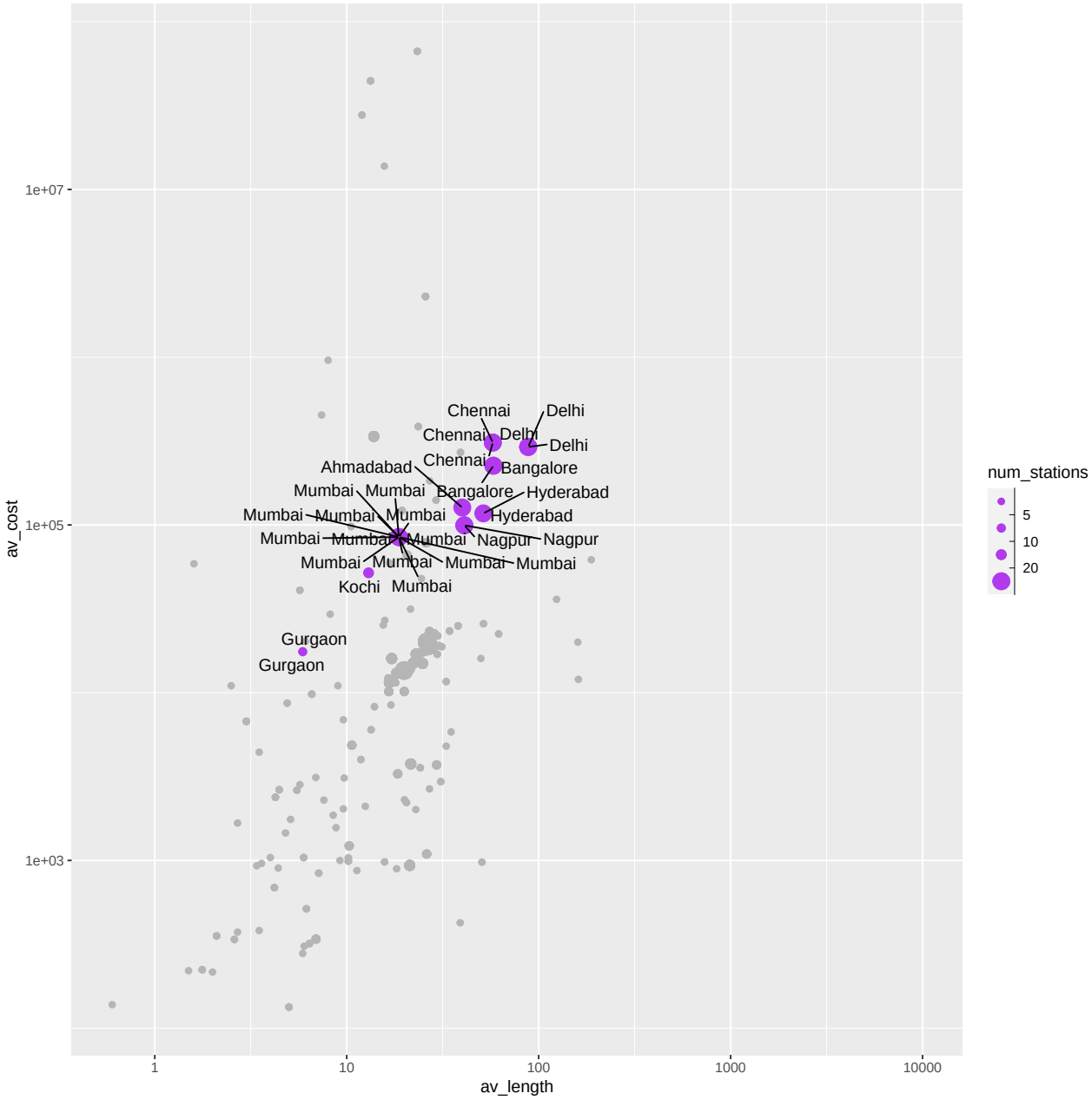
```
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations, color = india_y_n)) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20)) +
  scale_color_manual (values = c("darkgray", "darkorchid"))
```



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

```
stations_df <- data.frame(table(transit_cost$city)) %>% rename(num_stations = Freq, city = Var1)

d4 <- left_join(d3, stations_df, by = "city")

d4.0 <- d4 %>%
  group_by(stations) %>%
 filter(country  == "IN")


ggplot(d4,aes(av_length, av_cost)) +
  geom_point(aes(size = num_stations), color = "gray71") +
  geom_point(data = d4.0, aes(size = stations), color = "darkorchid2") +
```

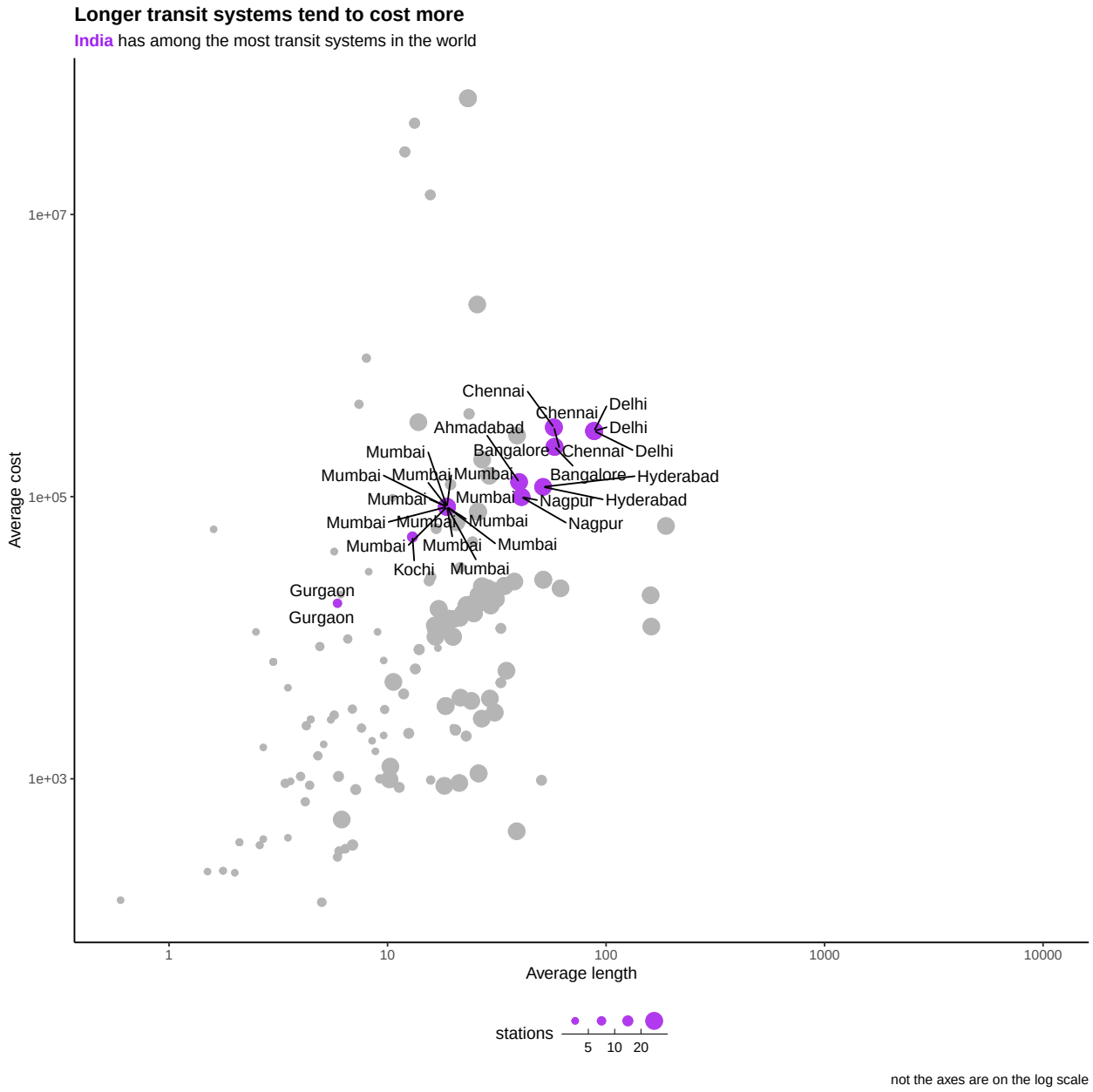
```
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20))+
geom_text_repel(data = d3.0, aes(label = city), max.overlaps = Inf)
```



8. Adjust the theme settings.

```
library(ggtext)
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
  geom_point(data = d3.0, aes(size = stations), color = "darkorchid2") +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20))+
```

```
geom_text_repel(data = d3.0, aes(label = city), max.overlaps = Inf) + theme_classic() +
  labs(title = "Longer transit systems tend to cost more",
       subtitle = "<b style='color:purple'>India</b> has among the most transit systems in the world", 
```

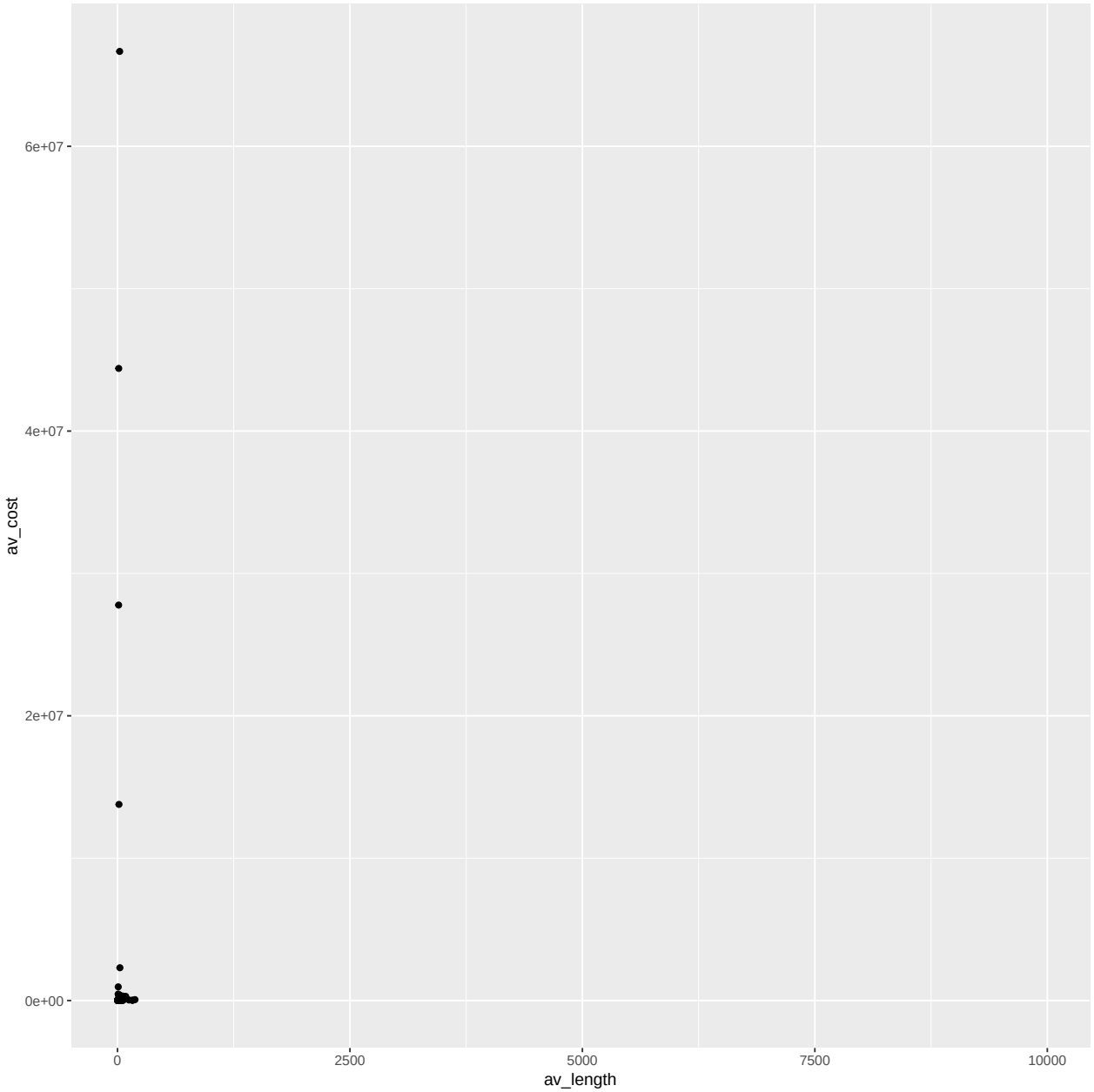


Question 2

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

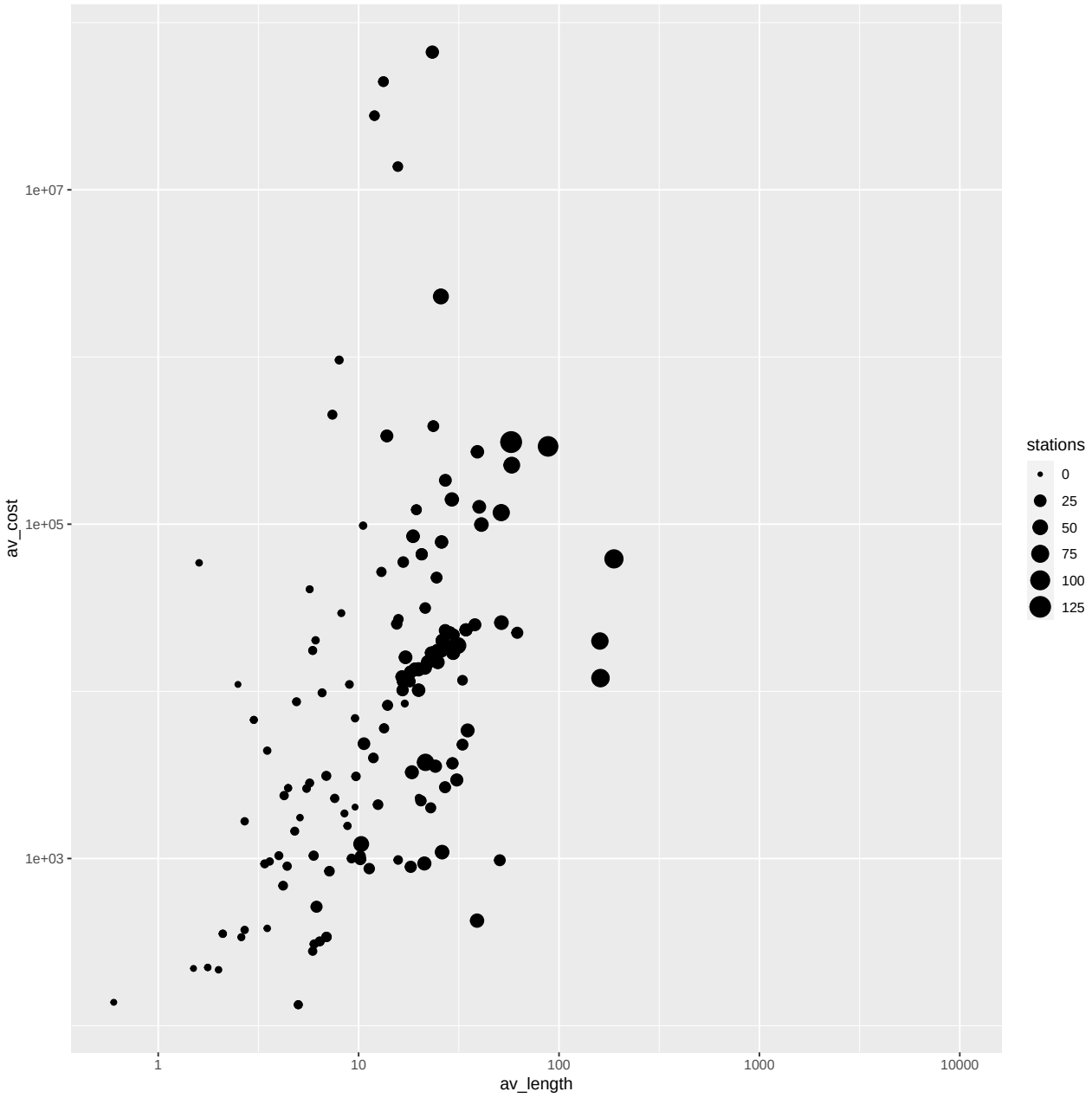
1. Compute the average length and average cost of transit systems by country and city.
2. Create a basic scatter plot by placing **Average Length** on the x-axis and **Average Cost** on the y-axis.

```
transit_cost %>%
group_by(country, city) %>%
summarize(av_length = mean(length, na.rm = T), av_cost = mean(cost, na.rm = T)) %>%
  ggplot(aes(av_length, av_cost)) + geom_point()
```



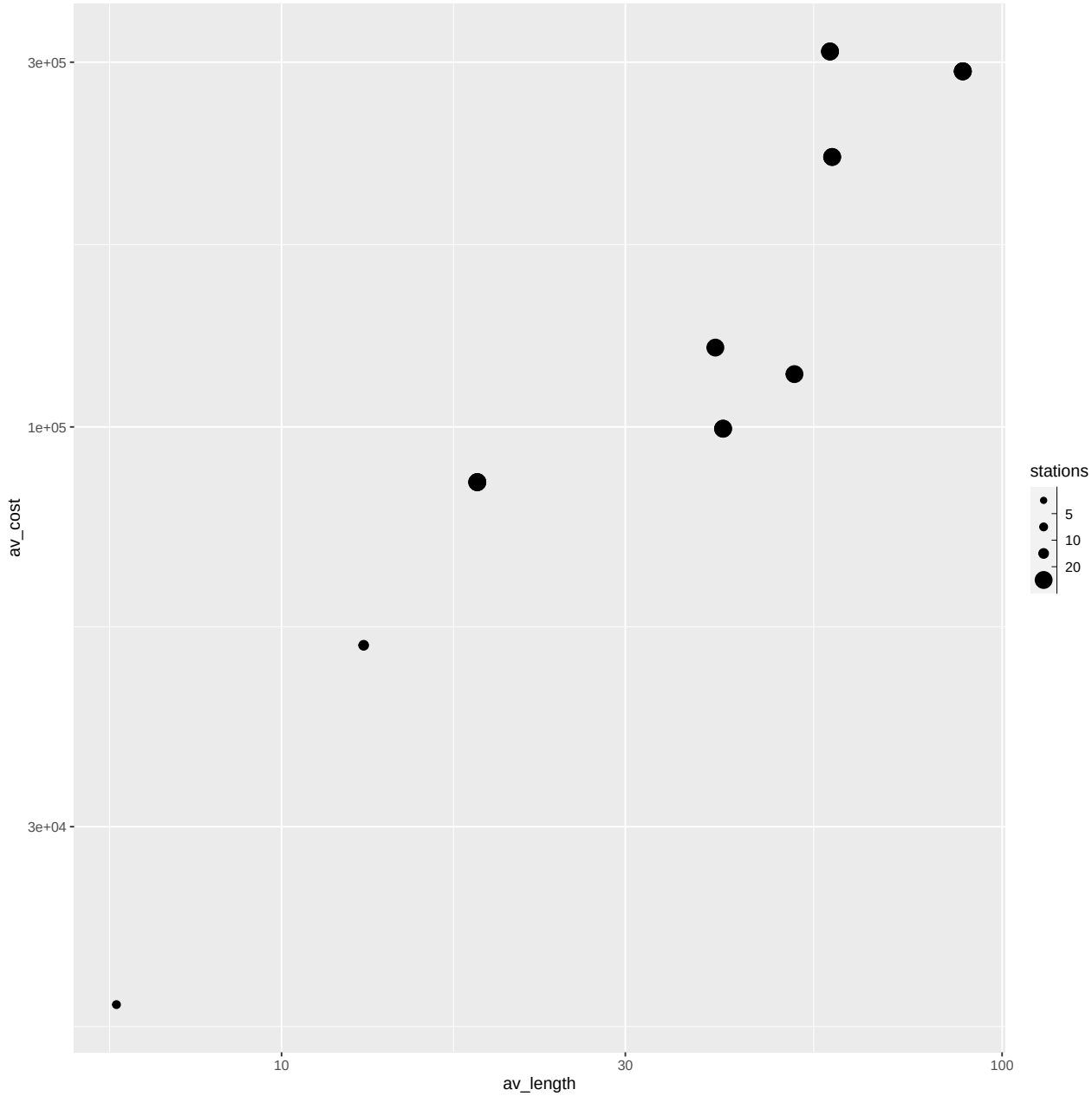
3. In the scatter plot,make the size of the data points represent the number of transit systems in that particular city (Hint: use `aes(size=)` within the `geom_point()` function).

```
d2 <- transit_cost %>%
group_by(city) %>%
summarize(av_length = mean(length, na.rm = T),
          av_cost = mean(cost, na.rm = T), stations)



ggplot(d2,aes(av_length, av_cost)) +
  geom_point(aes(size = stations)) +
  scale_x_log10() +
  scale_y_log10()
```



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()`)

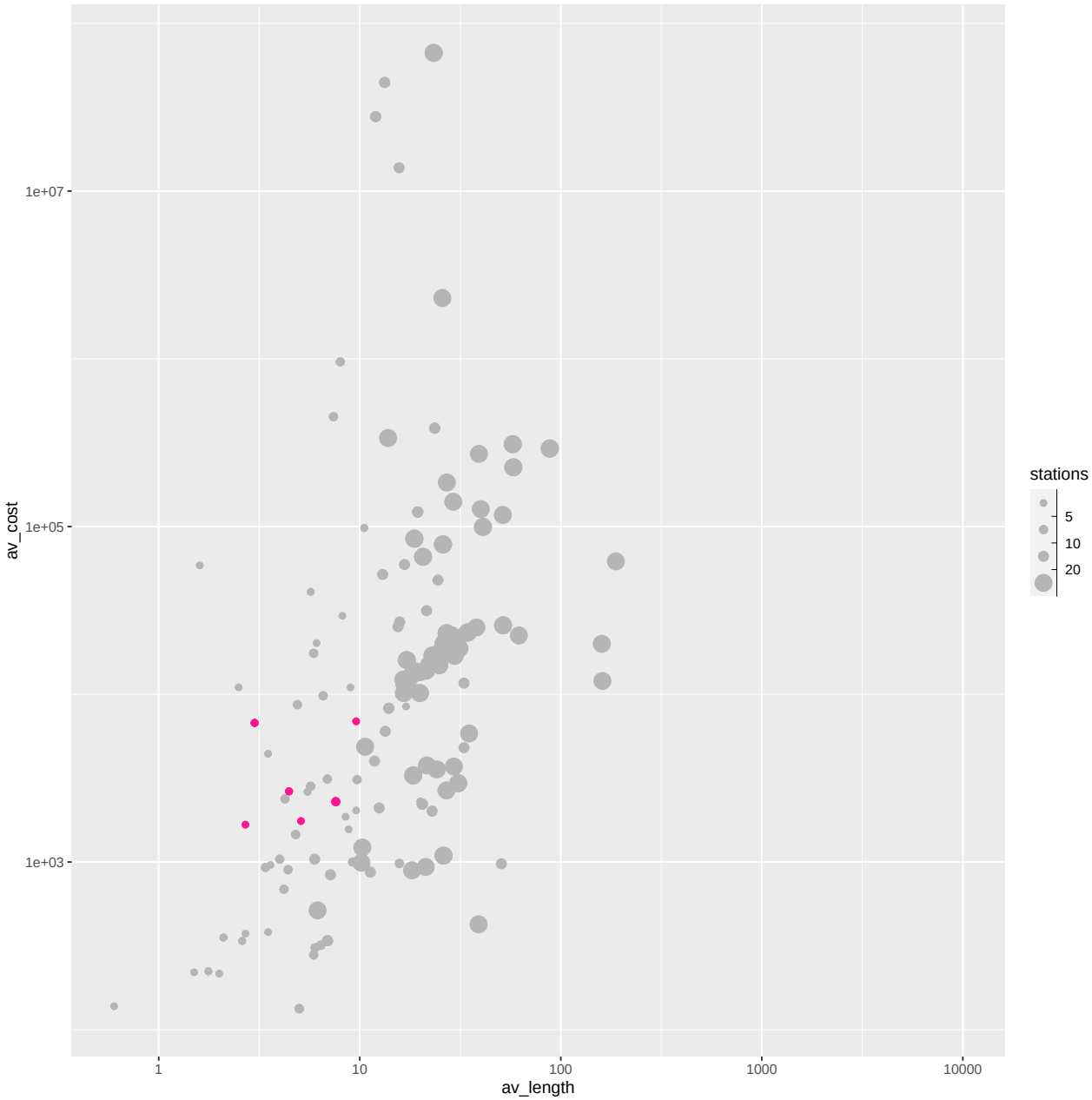
```
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(data = d3.0, aes(size = stations)) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20))
```



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.

```
d3.1 <- d3 %>%
    group_by(stations) %>%
    filter(country  == "US") %>%
    mutate(name = ifelse(country == "US", "United States", ""))
```

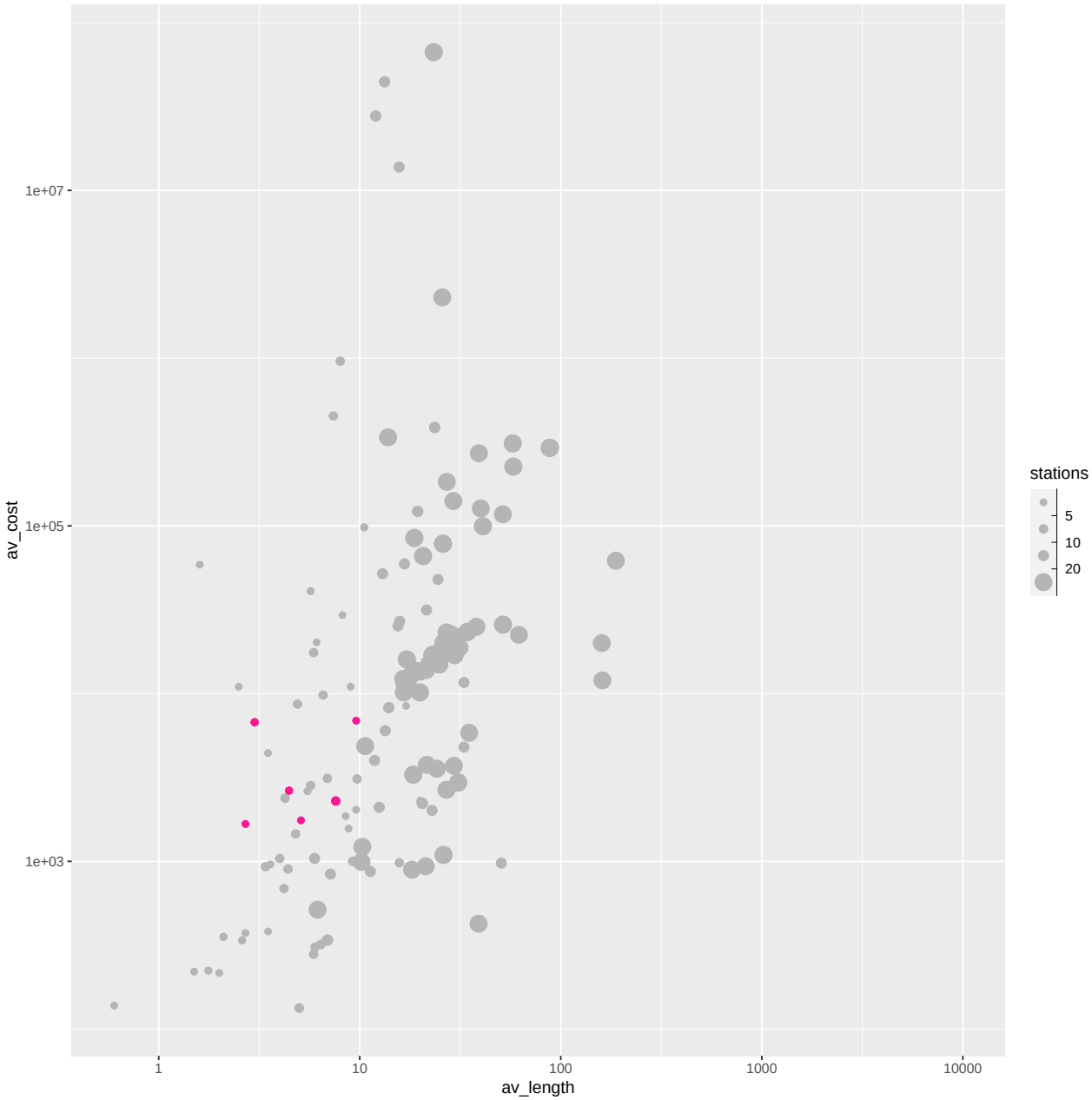
```
p2 <- ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
  geom_point(data = d3.1, aes(size = stations), color = "deeppink1", show_guide=FALSE) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20))


p2
```




```
#p2 +guides(color = FALSE)
#cornflowerblue
```

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.

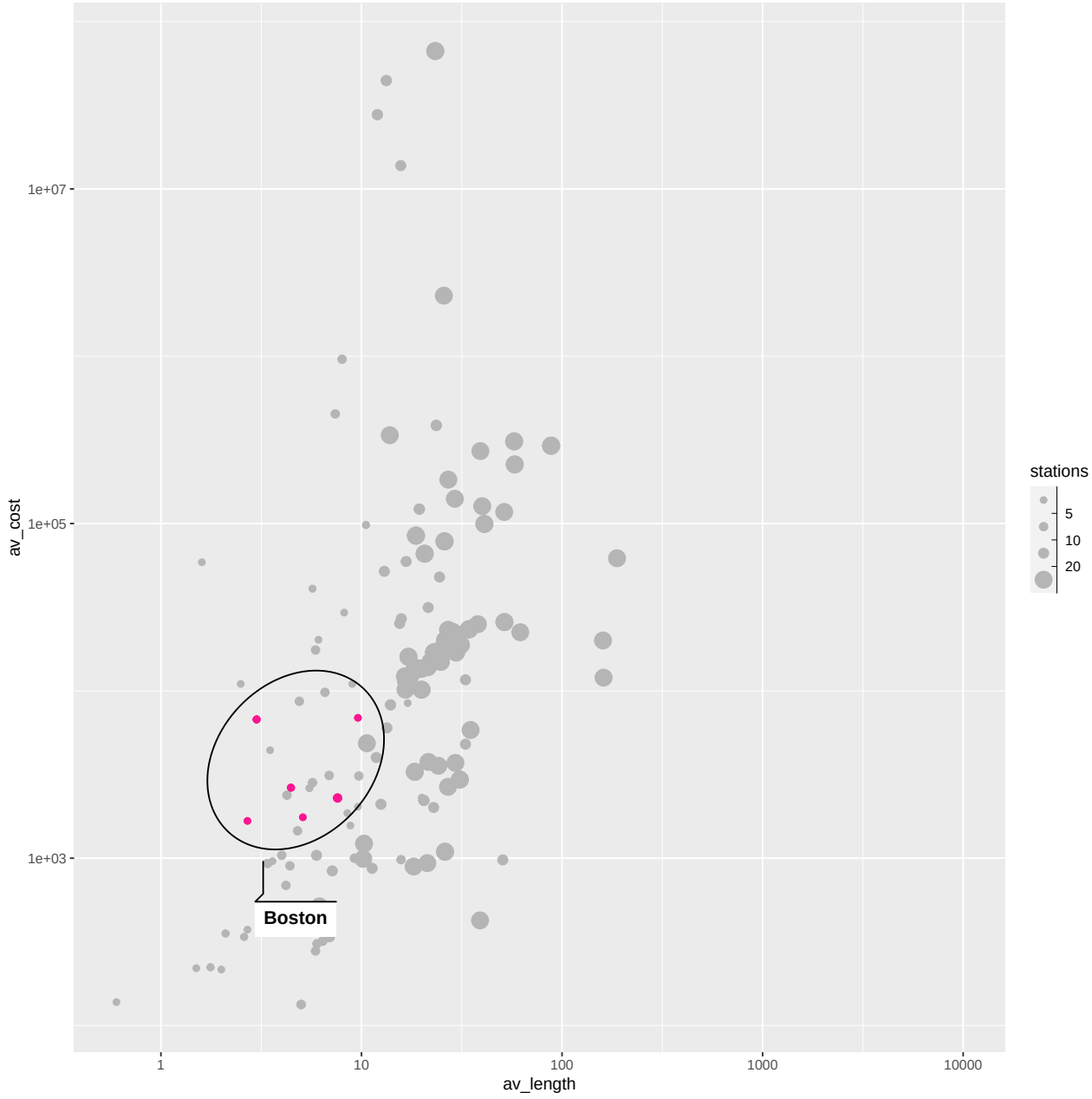
```
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
  geom_point(data = d3.1, aes(size = stations), color = "deeppink1", show_guide=FALSE) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20))
```



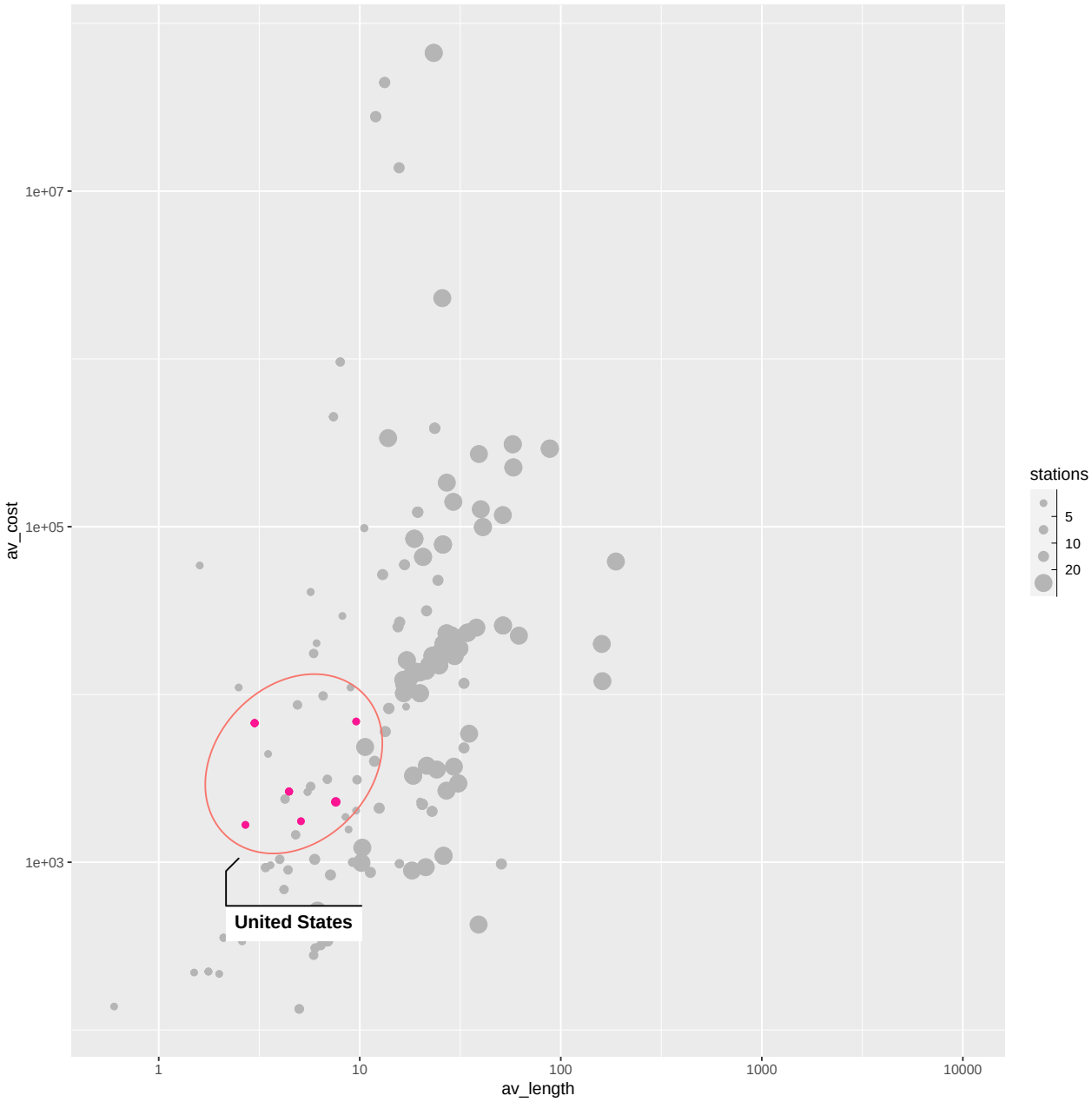
7. Using the `geom_mark_ellipse()` function from the `ggforce` package, circle the data points for the US cities.

```
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
```

```
  geom_point(data = d3.1, aes(size = stations), color = "deeppink1", show_guide=FALSE) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20)) +
  geom_mark_ellipse(data = d3.1, aes(label = city))
```

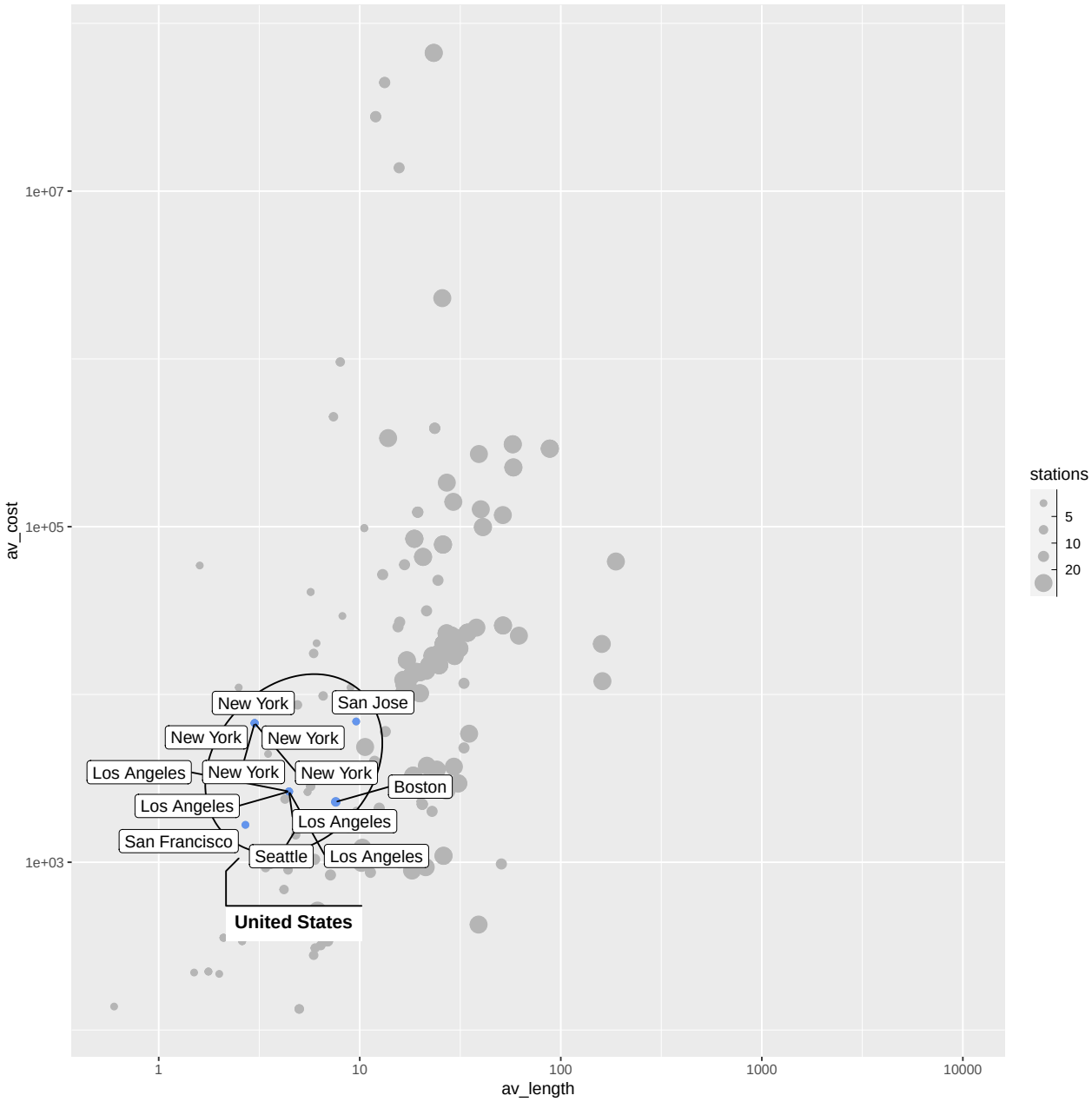



```
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
  geom_point(data = d3.1, aes(size = stations), color = "deeppink1", show_guide=FALSE) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20)) +
  geom_mark_ellipse(data = d3.1, aes(label = name, color = name), show_guide=FALSE) + scale_linetype_ma
```



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

```
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
  geom_point(data = d3.1, aes(size = stations), color = "cornflowerblue", show_guide=FALSE) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20)) +
  geom_mark_ellipse(data = d3.1, aes(label = name), show_guide=FALSE) +
  geom_label_repel(data = d3.1, aes(label = city), max.overlaps = Inf)
```

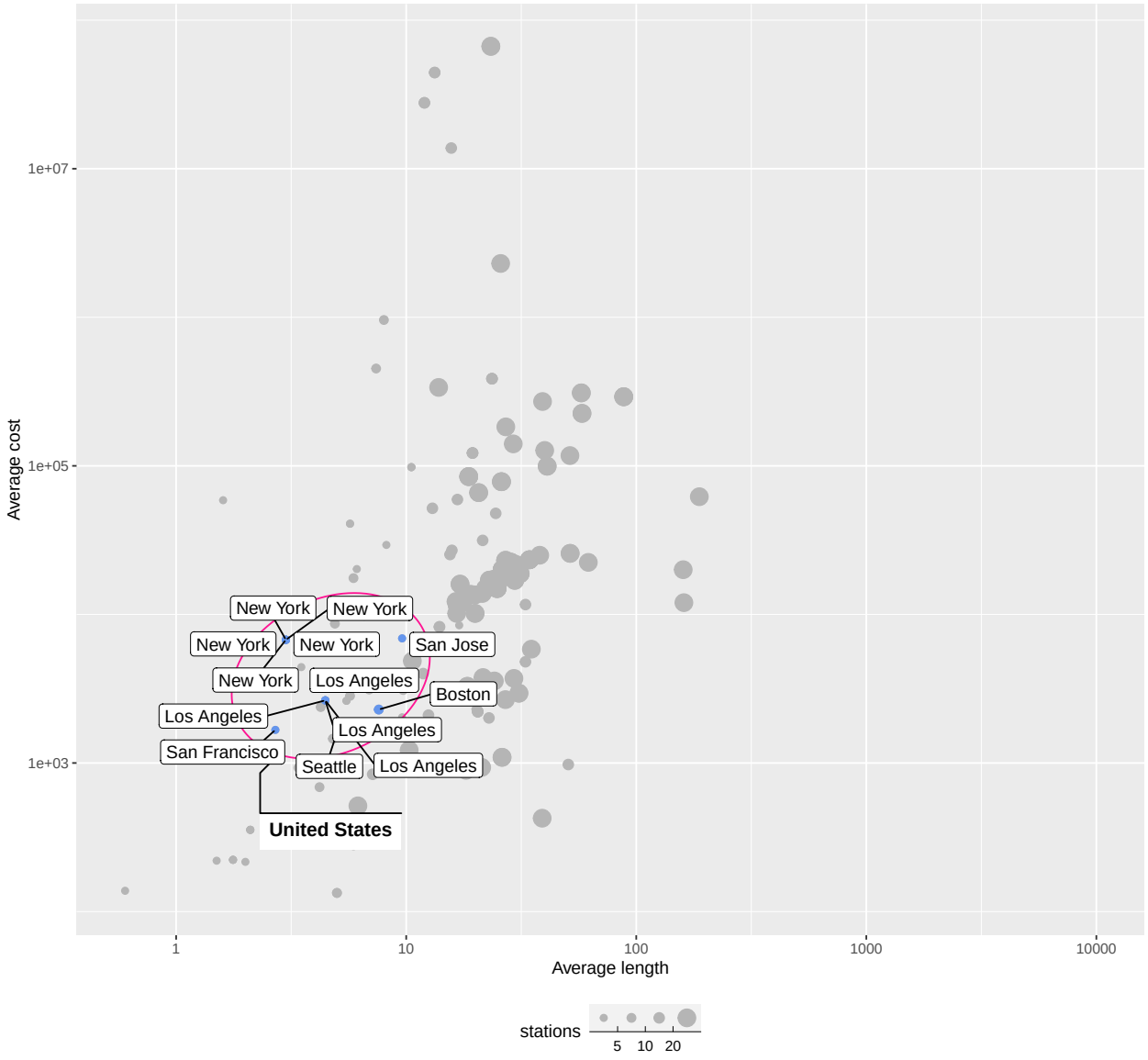


9. Adjust the theme settings.

```
ggplot(d3,aes(av_length, av_cost)) +
  geom_point(aes(size = stations), color = "gray71") +
  geom_point(data = d3.1, aes(size = stations), color = "cornflowerblue", show_guide=FALSE) +
  scale_x_log10() +
  scale_y_log10() +
  scale_size_binned(breaks =c(5,10,20)) +
  geom_mark_ellipse(data = d3.1, aes(label = name, color = name),color = "deeppink1",alpha = 0.05,  shou
  geom_label_repel(data = d3.1, aes(label = city, font = name), max.overlaps = Inf) +
labs(title = "Longer transit systems tend to cost more", subtitle = "United States has the most expensiv
```

Longer transit systems tend to cost more

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



not the axes are on the log scale