HW 1

Havi Khurana Cassie Malcom Merly Klaas

2/10/2022

Contents


```
#load packages
pacman::p_load(tidyverse, tidytuesdayR, dpylr, ggplot2, gghighlight, ggtext, ggrepel, rio, here, colorsp

transit_cost <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
    filter(!is.na(country)) #end rows have summary statistics
```

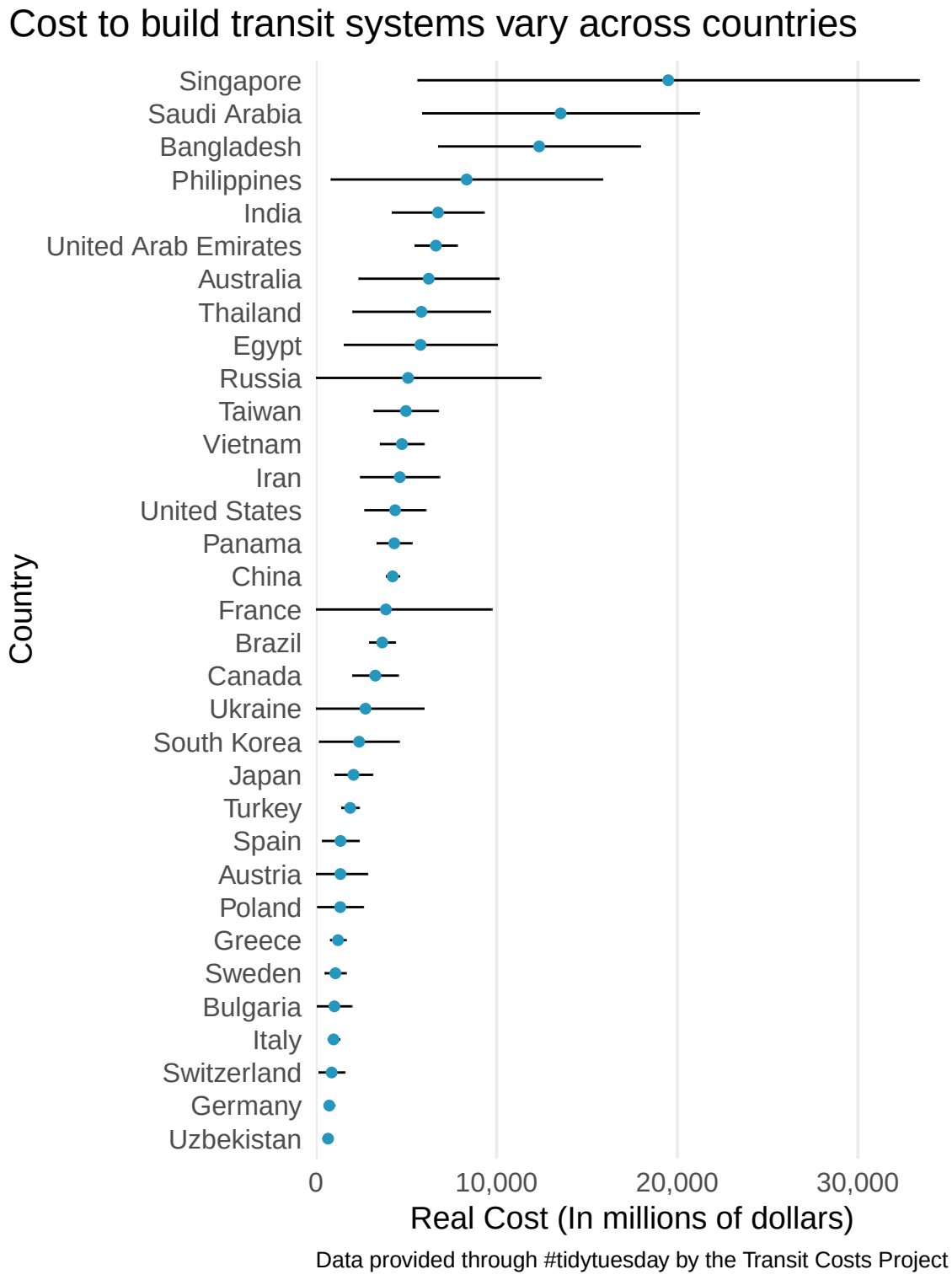
Question 1

```
#Data Cleaning

country_codes <- countrycode::codelist %>%
    select(country_name = country.name.en, country = ecb)

tc <- transit_cost %>%
        mutate(start_year = as.integer(start_year),
               end_year = as.integer(end_year),
               real_cost = as.numeric(real_cost)) %>%
    rename(id = e) %>% #didn't like naming convention
    group_by(country) %>% #adding means and standard errors for real_cost variable
    summarise(
        mean_rc = mean(real_cost),
        n=n(),
        sd=sd(real_cost),
        se=sd/sqrt(n)) %>%
    filter(n >= 3) %>%  #Filter out any country with less than 3 observations
    merge(country_codes, by = "country") #merging with country names
```

Use the transit costs data to reproduce the following plot.

```
p1 <- tc %>%
ggplot(aes(x = mean_rc, y = reorder(country_name, mean_rc))) +
  geom_errorbar(aes(xmin = ifelse(mean_rc - 1.96*se<0,0,mean_rc - 1.96*se),
                    xmax = mean_rc + 1.96*se,
                    width = 0)) +
```

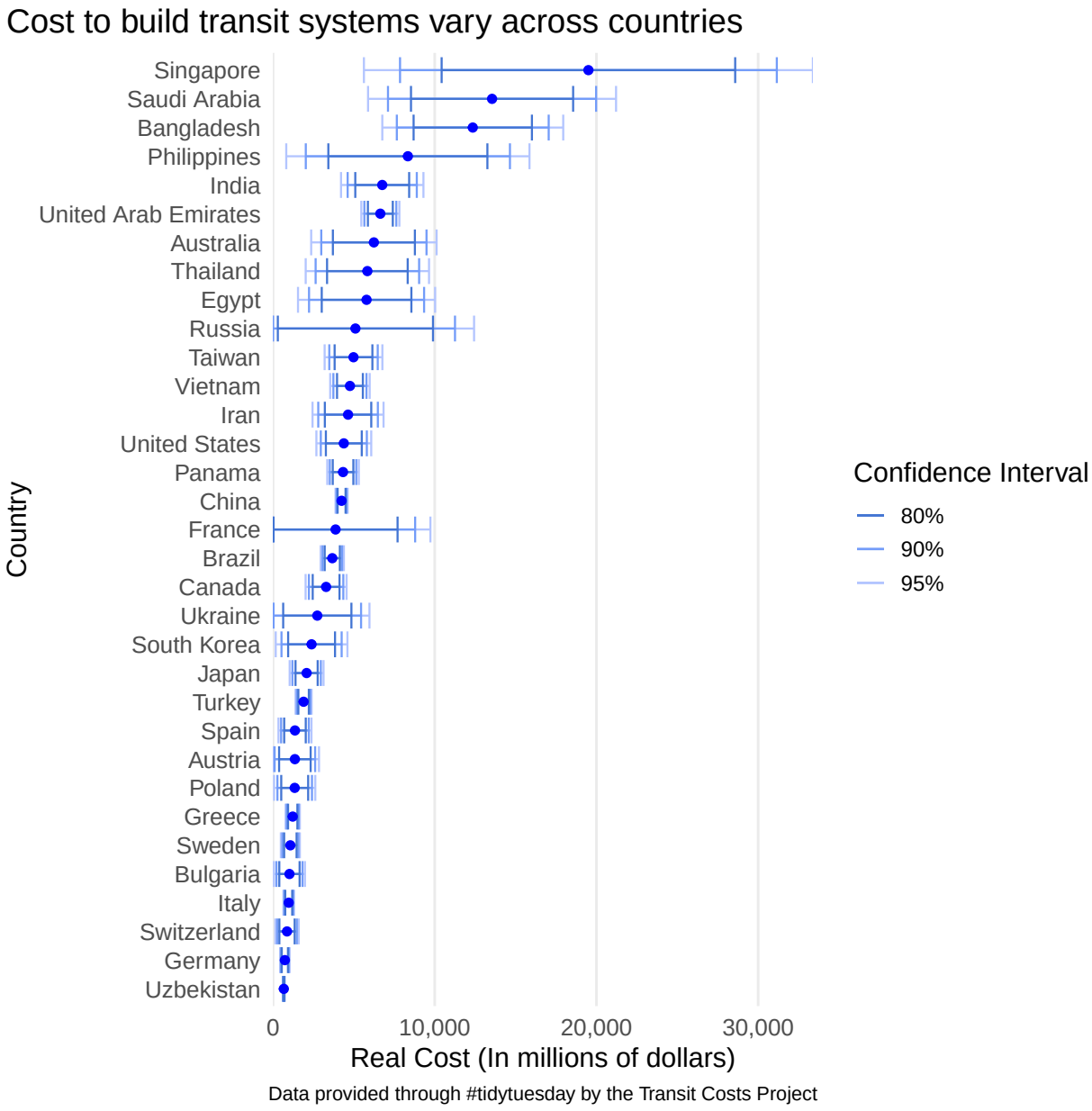
```
  geom_point(size = 1.8, colour = "#2596be") +
  theme_minimal(base_size = 14) +
  theme(plot.title.position = "plot", # easiest way to left align title
        plot.caption = element_text(hjust = 0.5, size = 10),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.x = element_blank(),
        axis.text = element_text(size = 12)) +
  labs(title = "Cost to build transit systems vary across countries",
       caption = "Data provided through #tidytuesday by the Transit Costs Project") +
  scale_x_continuous(name ="Real Cost (In millions of dollars)",
                     limits = c(0, 35000),
                     breaks = seq(0, 30000, 10000),
                     labels=c("0","10,000","20,000", "30,000"),
                     expand = c(0,0)) +
    scale_y_discrete(name = "Country")

p1
```



Question 2

Visualize the same relation, but displaying the uncertainty using an alternative method of your choosing - Multiple error bars.

```r
p2 <- tc %>%
ggplot(aes(x = mean_rc, y = reorder(country_name, mean_rc))) +
  geom_errorbar(aes(xmin = ifelse(mean_rc + qnorm(.025)*se<0,0,mean_rc + qnorm(.025)*se),
                    xmax = mean_rc + qnorm(.975)*se,
                color = "95%")) +
  geom_errorbar(aes(xmin = ifelse(mean_rc + qnorm(.05)*se<0,0,mean_rc + qnorm(.05)*se),
                    xmax = mean_rc + qnorm(.95)*se,
                color = "90%")) +
  geom_errorbar(aes(xmin = ifelse(mean_rc + qnorm(.1)*se<0,0,mean_rc + qnorm(.1)*se),
                    xmax = mean_rc + qnorm(.9)*se,
                color = "80%"))+
  geom_point(size = 1.8, colour = "blue") +
  scale_color_manual("Confidence Interval",
                     values = c("#4375D3", lighten("#4375D3", .3), lighten("#4375D3", .6))) +
  theme_minimal(base_size = 14) +
  theme(plot.title.position = "plot", # easiest way to left align title
        plot.caption = element_text(hjust = 0.5, size = 10),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.x = element_blank(),
        axis.text = element_text(size = 12)) +
  labs(caption = "Data provided through #tidytuesday by the Transit Costs Project",
       title = "Cost to build transit systems vary across countries") +
    scale_x_continuous(name ="Real Cost (In millions of dollars)",
                       breaks = seq(0, 30000, 10000),
                       labels=c("0","10,000","20,000", "30,000"),
                       expand = c(0,0)) +
    scale_y_discrete(name = "Country")

p2
```

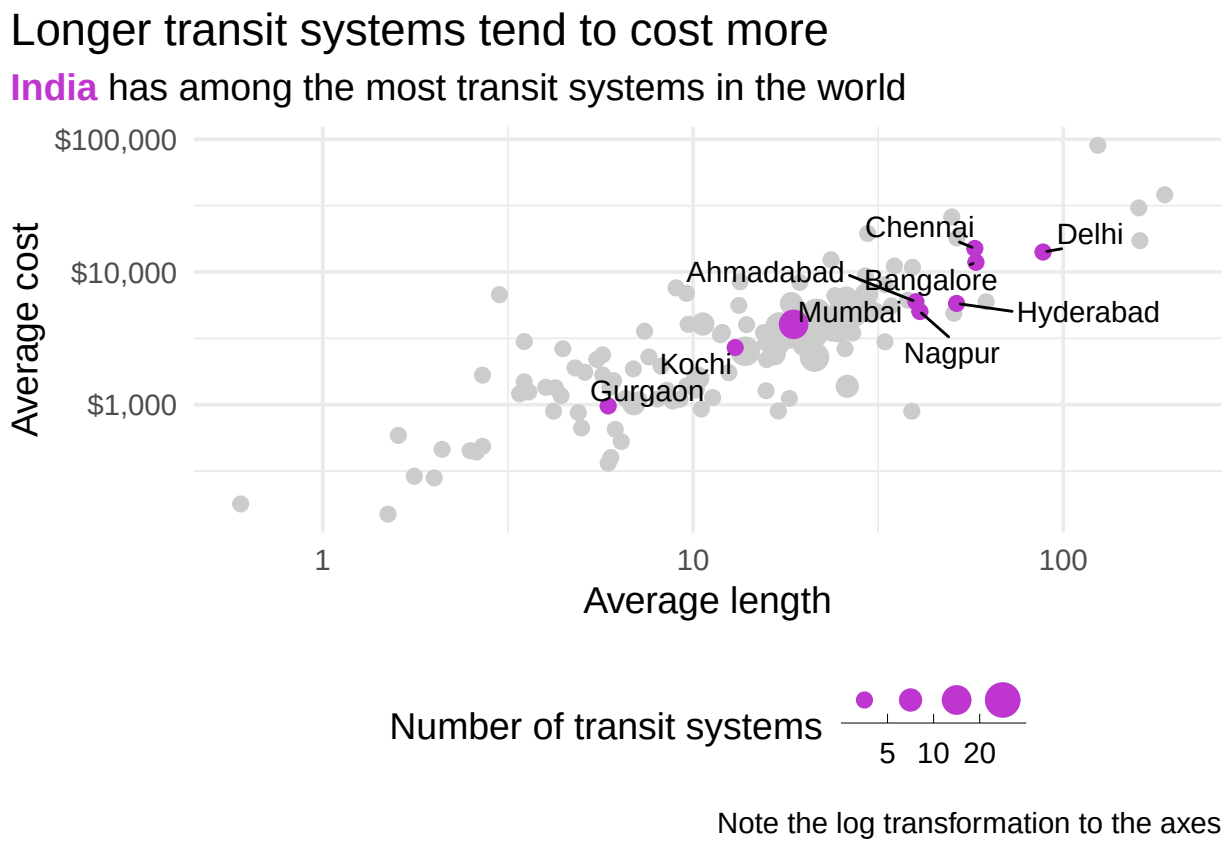


Question 3

Compute the mean length and real_cost by city. Reproduce the following plot.

```
transit_cost %>%
    group_by(country,city) %>%
    summarise(
        n = n(),
        length_mean = mean(length, na.rm = TRUE),
        real_cost_mean = mean(as.numeric(real_cost), na.rm = TRUE) #real_cost is char
    ) %>%
    ggplot(aes(x = length_mean, y = real_cost_mean))+
    geom_point(aes(size = n), color = "#bf35cf")+
```

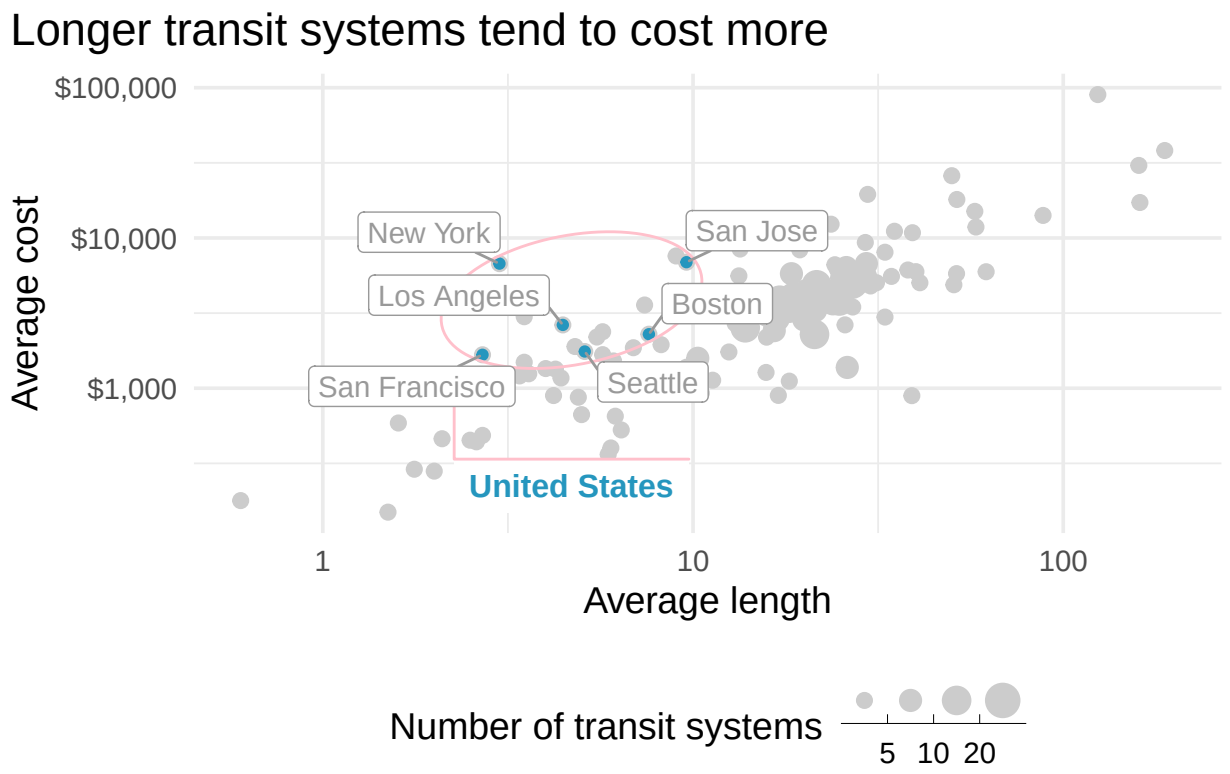
```
  scale_x_log10(limits = c(-1, 200))+
  scale_y_log10(labels = scales::dollar)+
  scale_size_binned(name = "Number of transit systems", breaks = c(5, 10, 20))+
  gghighlight(country == "IN",
              unhighlighted_params = list(color = "gray80"))+
  geom_text_repel(aes(label = city),
                  min.segment.length = 0)+
labs(
  title = "Longer transit systems tend to cost more",
  subtitle = "<span style = 'color: #bf35cf'>**India**</span> has among the most transit systems in t
  x = "Average length",
  y = "Average cost",
  caption = "Note the log transformation to the axes"
   )+
  theme_minimal(base_size = 14)+
  theme(
      plot.subtitle = element_markdown(),
      legend.position="bottom",
      plot.title.position = "plot"
  )
```



Question 4

Using basically the same data, reproduce the following plot. Note you'll need the country_name column in your dataset.

```r
q4 <- transit_cost %>%
    group_by(country,city) %>%
    summarise(
        n = n(),
        length_mean = mean(length, na.rm = TRUE),
        real_cost_mean = mean(as.numeric(real_cost), na.rm = TRUE) #real_cost is char
    ) %>%
    left_join(country_codes, by = "country")

q4 %>%
    ggplot(aes(x = length_mean, y = real_cost_mean))+
    geom_point(aes(size = n), color = "gray80")+
    scale_size_binned(name = "Number of transit systems", breaks = c(5, 10, 20))+
    geom_point(data = filter(q4,country == "US"), color = "#2596be", show.legend = FALSE)+
    scale_x_log10(limits = c(-1, 200))+
    scale_y_log10(labels = scales::dollar)+
    geom_mark_ellipse(aes(group = country,
                    label = country_name),
                    data = filter(drop_na(q4),
                                  country == "US"),
                    label.colour = "#2596be",
                    con.colour = "pink",
                    color = "pink",
                    expand = unit(1, "mm"),
                    con.type = "elbow")+
    geom_label_repel(data = filter(drop_na(q4),country == "US"),
                     aes(label = city),
                    min.segment.length = 0,
                    color = "gray60")+
  labs(
    title = "Longer transit systems tend to cost more",
    x = "Average length",
    y = "Average cost",
    caption = "Note the log transformation to the axes"
     )+
    theme_minimal(base_size = 14)+
    theme(
        plot.subtitle = element_markdown(),
        legend.position="bottom",
        plot.title.position = "plot"
    )
```



Question 5

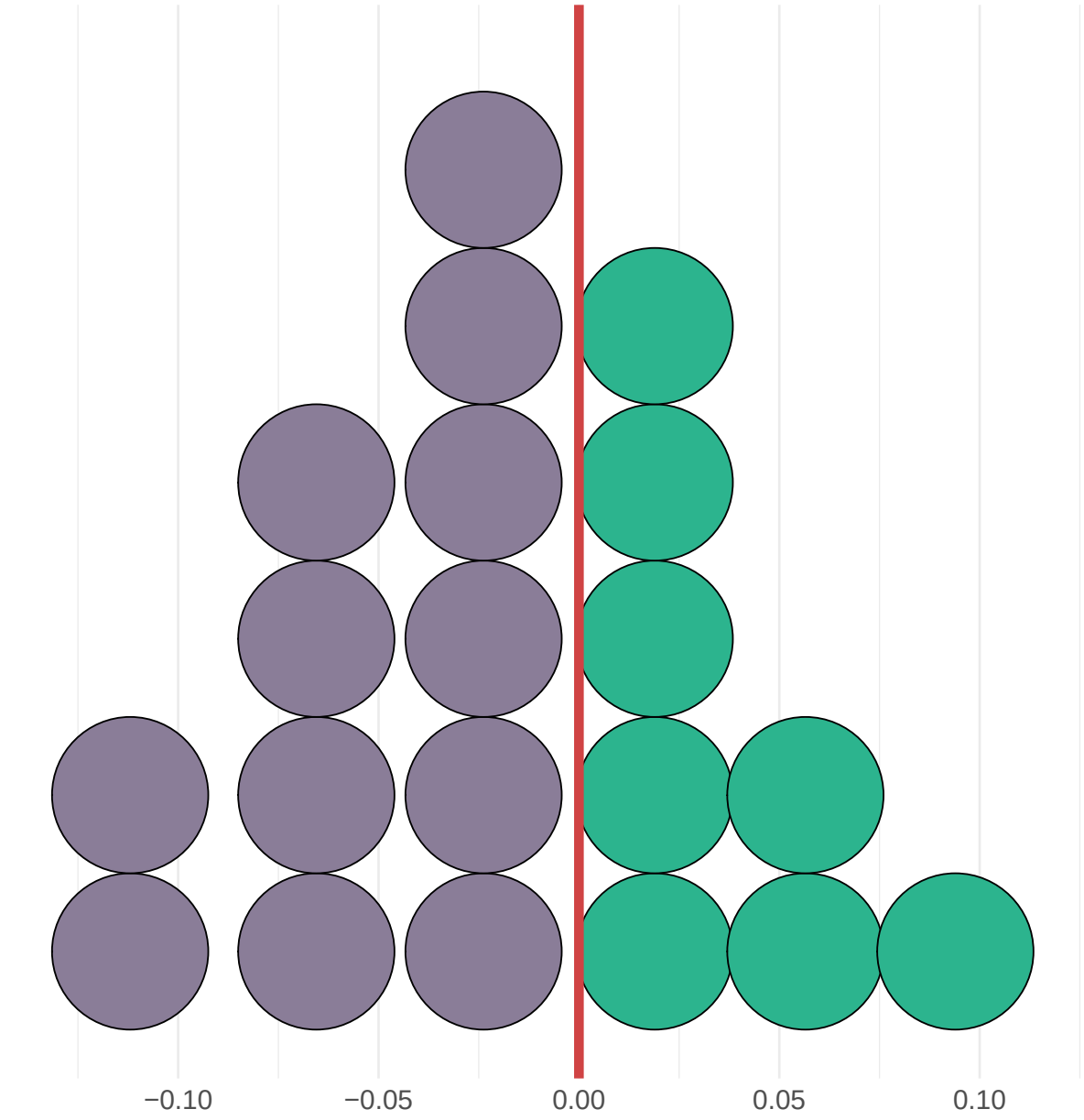
Use the crime dataset to run the following code and fit the corresponding model. Note, it may take a moment to run.

```
crime <- import(here("data", "crime.csv")) %>%
    janitor::clean_names()
```

```
model_data <- crime %>%
  mutate(neighborhood_id = relevel(factor(neighborhood_id), ref = "barnum"))

m <- glm(is_crime ~ neighborhood_id,
         data = model_data,
         family = "binomial")

# Extract the output using broom::tidy
tidied <- broom::tidy(m)

wbarnum <- tidied %>%
  filter(term == "neighborhood_idbarnum-west")

qnorm(ppoints(20),
      mean = wbarnum$estimate,
      sd = wbarnum$std.error)
```

```r
dizcretized <- data.frame(
    x = qnorm(ppoints(20),
      mean = wbarnum$estimate,
      sd = wbarnum$std.error)) %>%
        mutate( wbarnum = ifelse( x <= 0,"#8A7D98","#2CB48E" ))
```

```r
#Discretized plot
ggplot(dizcretized, aes(x)) +
  geom_dotplot(aes(fill = wbarnum), binwidth = 0.039) +
     geom_vline(xintercept = 0,
             color = "#D04344",
             linetype = "solid",
             size = 2) +
    scale_fill_identity(guide = "none") +
     scale_y_continuous(name = "",
                        breaks = NULL) +
    labs(title = "Probability of different crime rates between neighborhoods",
         subtitle = "<span style = 'color: #8A7D98'>**West Barnum**</span> compared to <span style = 'c
         x = "Difference in log odds of a crime being committed",
          caption = "Each ball represents 5% probability") +
    theme_minimal() +
     theme(plot.subtitle = element_markdown(),
           text = element_text(size=15))
```




```
#ggsave(here("plots", "Discretized-plot.pdf"))
```