HW 1

Havi Khurana Cassie Malcom Merly Klaas

2/10/2022

Contents


```
transit_cost <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
```

Question 1

Use the transit costs data to reproduce the following plot.

Question 2

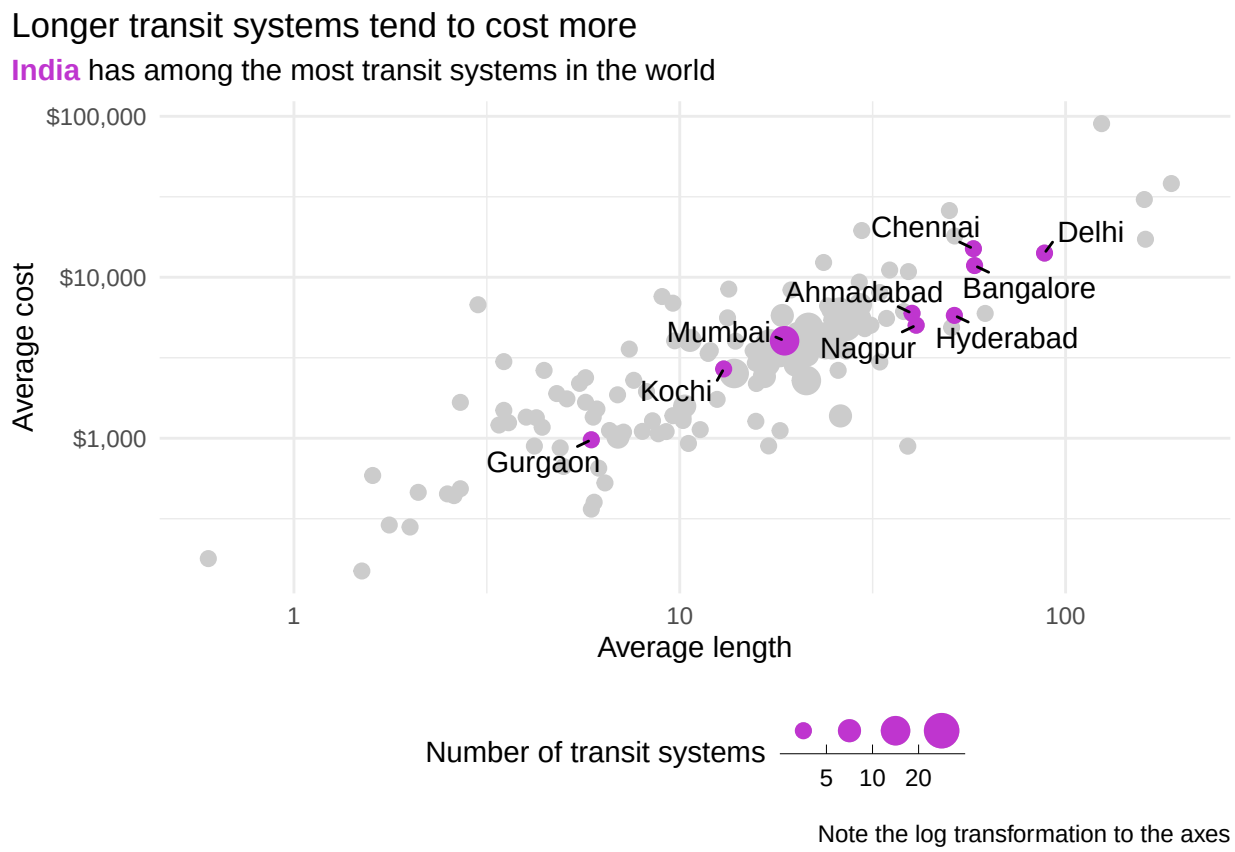
Visualize the same relation, but displaying the uncertainty using an alternative method of your choosing.

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 th
  x = "Average length",
  y = "Average cost",
  caption = "Note the log transformation to the axes"
   )+
  theme_minimal()+
  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.

Question 5

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

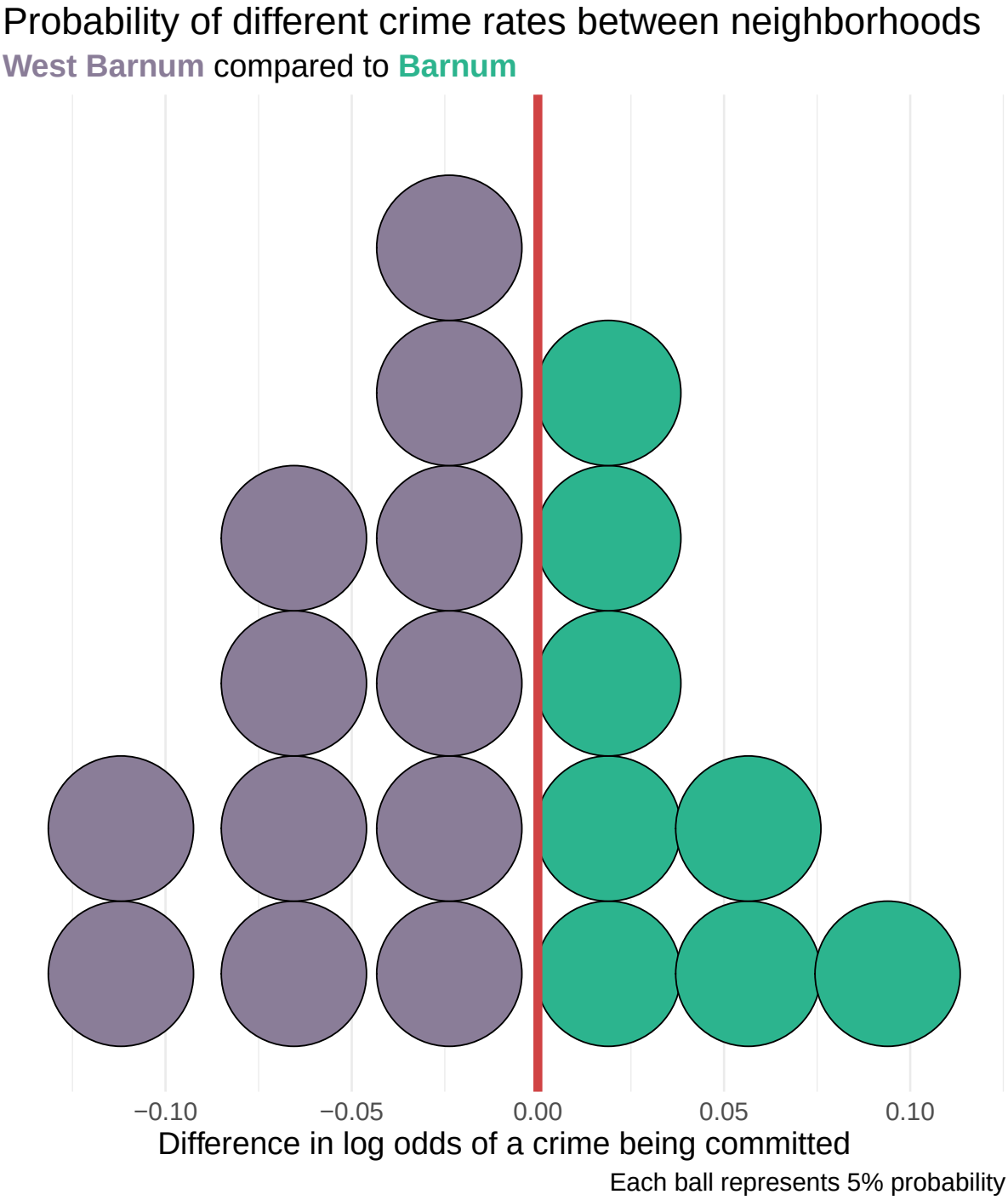
```
crime <- import(here("data", "crime.csv")) %>%
    janitor::clean_names()
```

```
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)
```

```
##  [1] -0.126627430 -0.097344692 -0.081073527 -0.068933536 -0.058852089
##  [6] -0.049981453 -0.041879243 -0.034276401 -0.026988755 -0.019876057
## [11] -0.012819518 -0.005706821  0.001580825  0.009183667  0.017285878
## [16]  0.026156513  0.036237960  0.048377952  0.064649117  0.093931854
```

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

```
#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"))
```