

Quiz 4

Task

Study Information.

The data was collected by the Sustainable Development Policy Institute (SDPI), which is a non-profit think tank in Pakistan. The data was made accessible on [Open Now Pakistan](#). [This Kaggle dataset](#) extracted the data from here, and merged all the data across districts so that each district was an observation in the file.

```
library(sf)
library(tidyverse)
library(MASS)
library(broom)
library(grid)
library(ggrepel)
library(knitr)
library(glmmTMB)
library(broom.mixed)
```

```
districts <- read_csv("data/Pakistan Districts Profile.csv")
```

```
districts <- districts |>
  mutate(doctors = `Number of Doctors`,
         poverty = as.numeric(`Multi Dimensional Poverty Index`),
         pop = `Population`,
         name = `name`,
         province = Province,
         public_hospitals = if_else(`Govt. Health Institutions` == 0, 0, 1))
```

```
districts <- dplyr::select(districts, doctors, poverty,
                           pop, name, province, public_hospitals)
```

i Introduction to the Data.

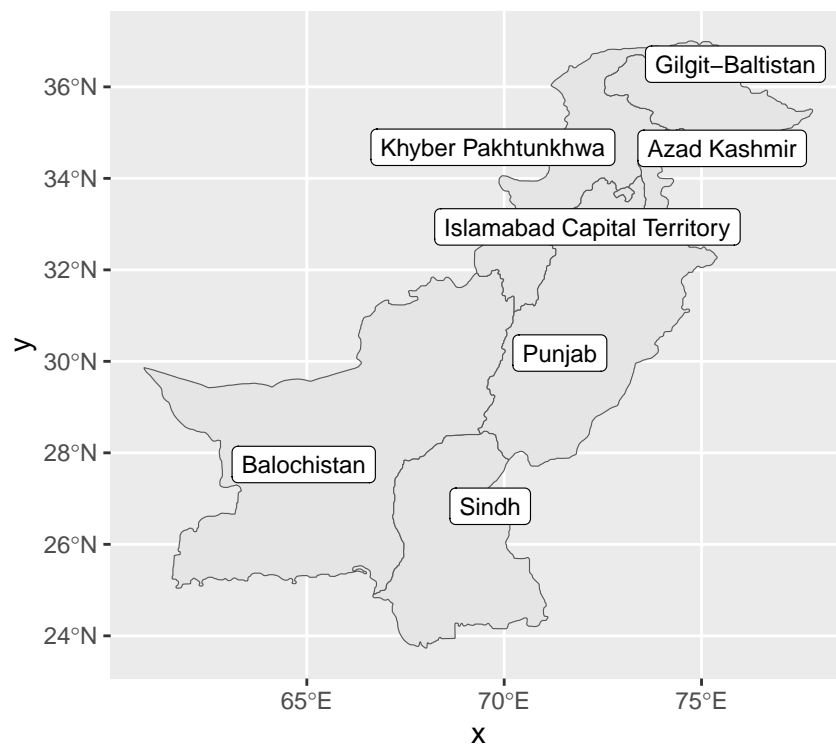
Your task is to model the number of doctors in Pakistan based on poverty levels.

- **districts:** Information about different geographic districts of Pakistan. This was taken from Kaggle (1), which extracted the data from Open Data Pakistan (2).
 - *doctors*, the number of doctors in a district. Our response variable.
 - *pop*, the size of the pop force.
 - *poverty*, a poverty index
 - *province*, a larger geographic region. Just like each state has multiple cities in the US, each province has multiple districts in Pakistan. This will be useful in visualizing how numbers vary across provinces.

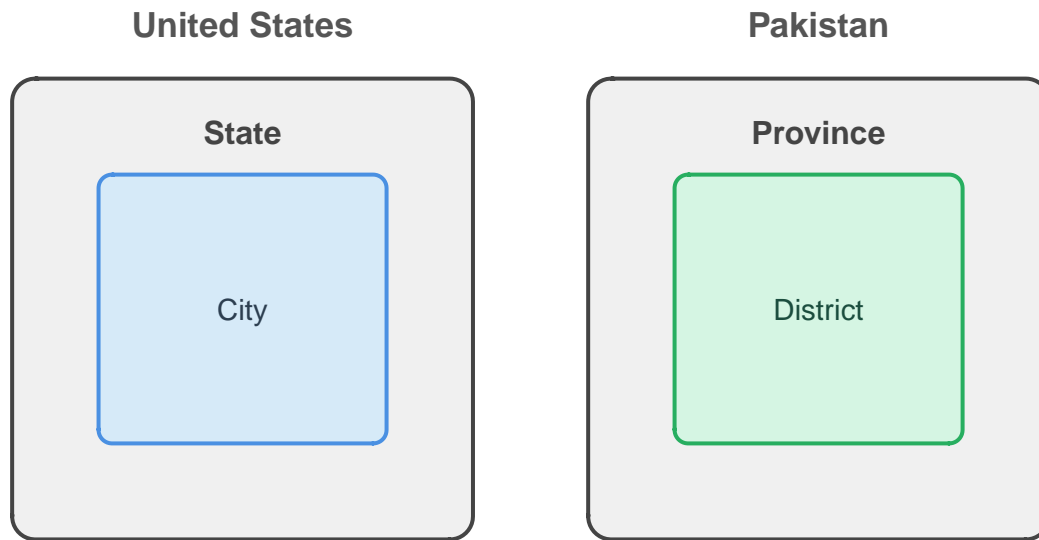
What is a district and province?

Terms for geographical divisions in Pakistan differ from those used in the United States. The United States has ~50 states; within each state, there are multiple cities. Similar to states, the Pakistani government refers to larger regions as provinces. While there are historically 4 official provinces in Pakistan, there are two additional regions with special status and the capital territory, so there are seven regions that can be considered provinces. For this analysis, we will consider there to be 7 provinces. In each province, there are multiple districts.

```
pakistan <- st_read("data/geoBoundaries-PAK-ADM1.shp", quiet = TRUE)
ggplot(pakistan) +
  geom_sf() +
  geom_label_repel(
    aes(label = shapeName, geometry = geometry),
    stat = "sf_coordinates", # computes centroids for label placement
    size = 3,
    box.padding = 0.3,
    label.size = 0.2,      # border thickness of white box
    label.r = unit(0.15, "lines"), # corner radius
    fill = "white",        # box background color
    color = "black"        # text color
  )
```



Geographical Hierarchy

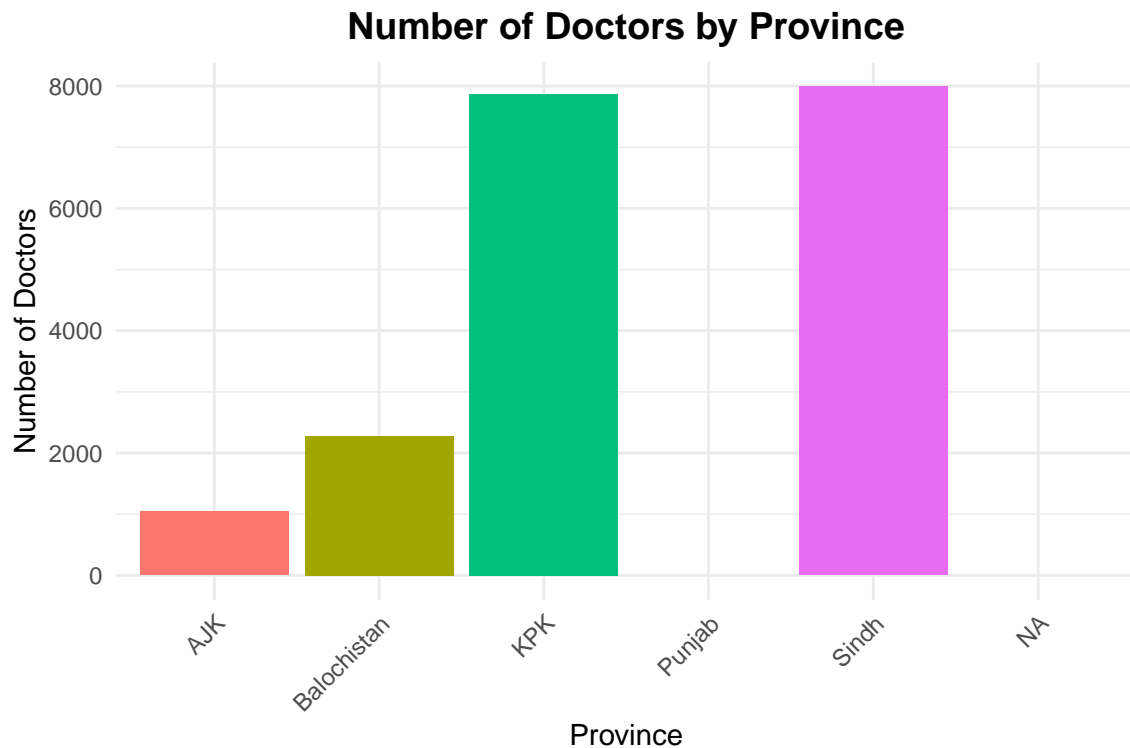


determined how to do grid nested boxes from this conversation: <https://chatgpt.com/c/6807c215-16e8-8008-a4ee-e24c79226624>

Question 1: Make a visualization highlighting number of doctors in each of Pakistan's 7 provinces. What do you notice? Optional: Feel free to find geographic identification data online, if you're interested in creating a map visual.

Students can simply do a bar plot with number of doctors versus province as shown below:

```
ggplot(districts, aes(x = province, y = doctors, fill = province)) +  
  geom_col(show.legend = FALSE) +  
  labs(  
    title = "Number of Doctors by Province",  
    x = "Province",  
    y = "Number of Doctors"  
  ) +  
  theme_minimal() +  
  theme(  
    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),  
    axis.text.x = element_text(angle = 45, hjust = 1)  
  )
```



Alternatively, here's another visualization students can make:

```
provinces <- districts |>
  group_by(province) |>
  summarise(
    doctors = sum(doctors)
  )
```

```
provinces <- provinces |>
  mutate(province = recode(province,
    "KPK" = "Khyber Pakhtunkhwa",
    "AJK" = "Azad Kashmir"))
```

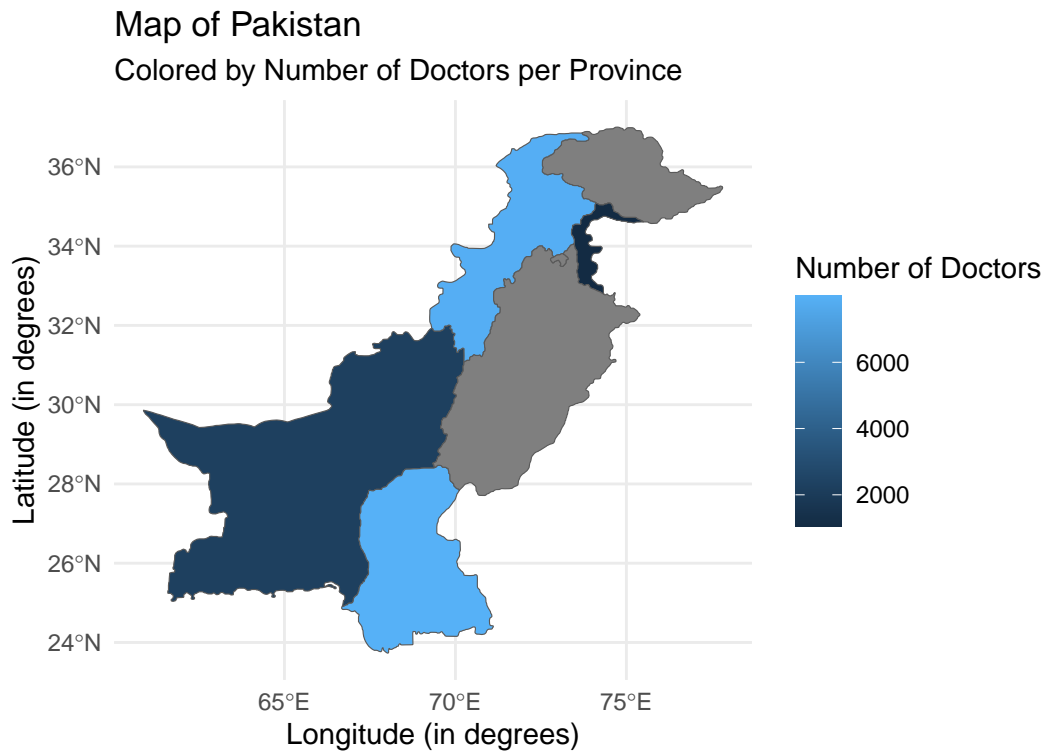
```
pak <- left_join(pakistan, provinces, by = c("shapeName" = "province"))
```

```
ggplot(pak) +
  geom_sf(aes(fill = doctors)) +
  theme_minimal() +
  labs(
    x = "Longitude (in degrees)",
```

```

y = "Latitude (in degrees)",
title = "Map of Pakistan",
subtitle = "Colored by Number of Doctors per Province",
fill = "Number of Doctors"
)

```



Things to notice:

- Missing data: Dark grey is the default when there is an NA for the number of doctors in a province. We don't have information for number of doctors in a couple of provinces. Not Punjab is dark grey as well as Gilgit Baltistan and Islamabad Capital Territory.
- Number of Doctors: The blue shading helps shed light into number of doctors with lighter blue representing more doctors. There are the highest number of doctors in KPK and Sindh.

Question 2: What kind of model may be well suited for the task? Why?

Poisson is a good option when looking for counts because it will be defined on non-negative integers, among other reasons.

i Additional Instructions.

Before proceeding, we will remove observations where data is missing for doctors, pop, and poverty. This involves assuming that the observations are missing at random, but this criteria may not hold. For the sake of this analysis, we will proceed like this. Please incorporate the below code chunk into your code.

```
districts_two <- districts
districts <- districts |>
  filter(!is.na(doctors)) |>
  filter(!is.na(pop)) |>
  filter(!is.na(poverty))
```

Question 3: Is there evidence of overdispersion? Check for it and if there is, suggest how we might correct for it.

```
mean(districts$doctors)
```

```
[1] 224.8701
```

```
var(districts$doctors)
```

```
[1] 80391.51
```

There is evidence of overdispersion since the empirical variance exceeds the mean.

To correct for this, it may be useful to either utilize negative binomial regression or inflate the standard errors.

Question 4: Could an offset be useful in our goal to analyze covariates related to number of doctors?

An offset is useful because more populous areas naturally have more doctors; thus, we use pop force size as the offset.

Question 5: Depending on your answer for question 2, fit a model using either the regular Poisson (without inflating standard errors) or the negative binomial. Make a choice to include an offset depending on question 4. Your model should incorporate the covariate poverty. If you do include an offset, remember that $\log(0)$ is undefined and this may mean some data filtering is needed beforehand.

```
loca_new <- districts |>
  filter(pop != 0)
```

```
model <- glm.nb(doctors ~ poverty + offset(log(pop)), data = loca_new)

tidy(model, conf.int = TRUE) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-8.037	0.161	-49.809	0.000	-8.347	-7.716
poverty	-1.403	0.432	-3.249	0.001	-2.250	-0.557

Question 6: Interpret the effect of poverty.

The p-value is low. For each increase in the poverty index by one unit, the expected rate of doctors per pop population in a district will change by a multiplicative factor of 0.2143811. This means we expect the number of doctors to decrease by 78.6% for each point increase in poverty index.

There are no poverty index values with a one point difference since the poverty index appears to occur between 0 and 1. It may be more meaningful to make interpretations for each 0.1 point increase in the poverty index. For example, the expected rate of doctors per pop population in a district will change by a multiplicative factor of 0.857272. This means we expect the number of doctors to decrease by 14.3% for each 0.1 point increase in poverty index.

```
min(loca_new$poverty)
```

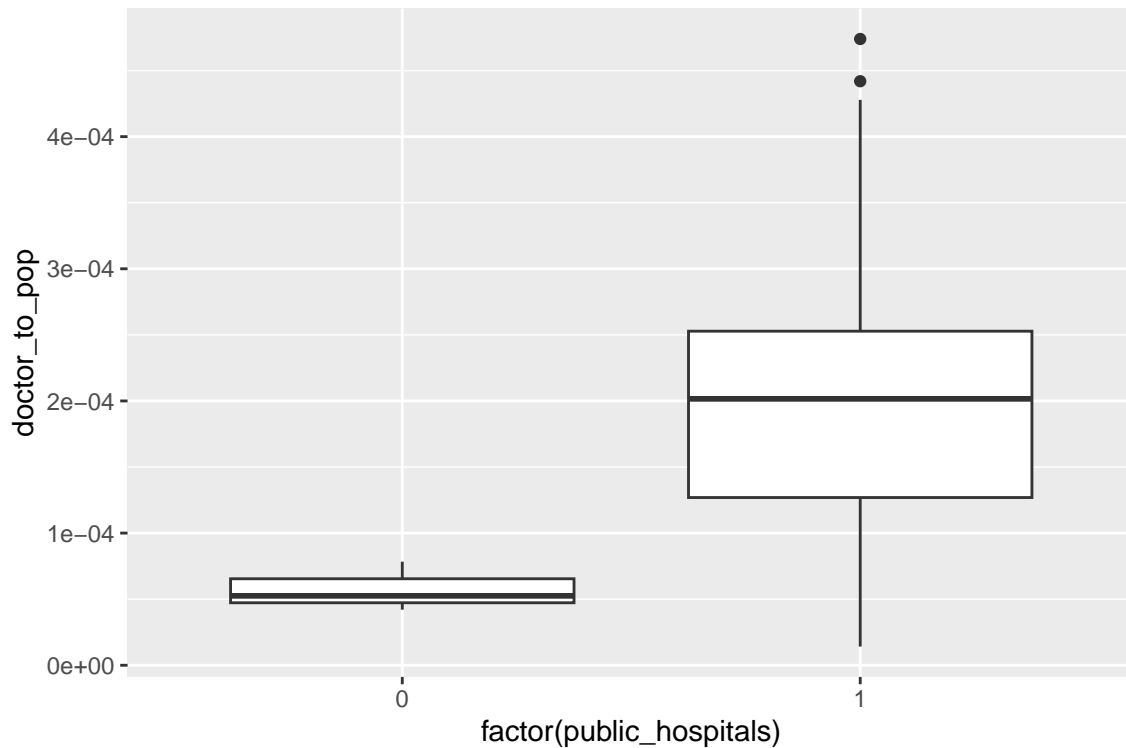
```
[1] 0.019
```

```
max(loca_new$poverty)
```

```
[1] 0.641
```

Question 7: Create a visualization and summary table to verify or reject the following statement: Districts with public hospitals in Pakistan have more similar doctor-to-pop force ratios than those without public hospitals

```
districts |>
  mutate(
    doctor_to_pop = doctors/pop
  ) |>
ggplot(aes(x = factor(public_hospitals), y = doctor_to_pop)) +
  geom_boxplot()
```

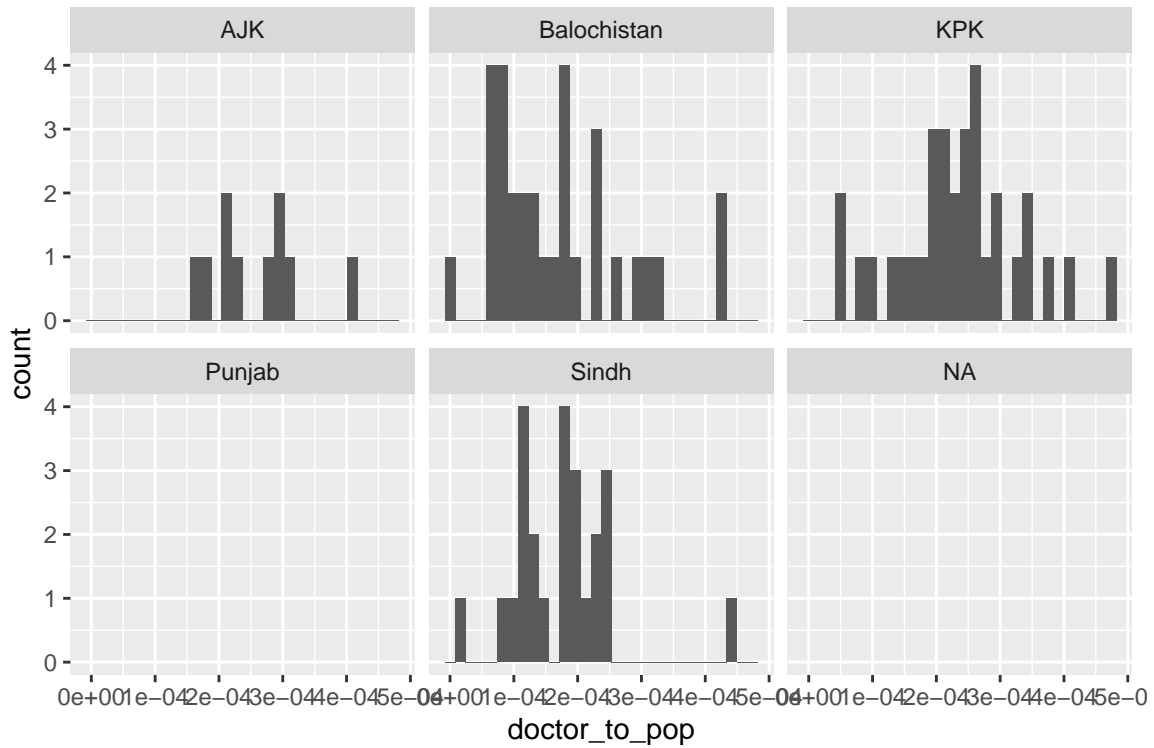


Question 7: Create a visualization and summary table to verify or reject the following statement: Districts in the same province in Pakistan have more similar doctor-to-pop force ratios than those in different provinces.

```
districts_two |>
  mutate(
    doctor_to_pop = doctors/pop
  ) |>
  ggplot(aes(x = doctor_to_pop)) +
  geom_histogram() +
  facet_wrap(~province)
```

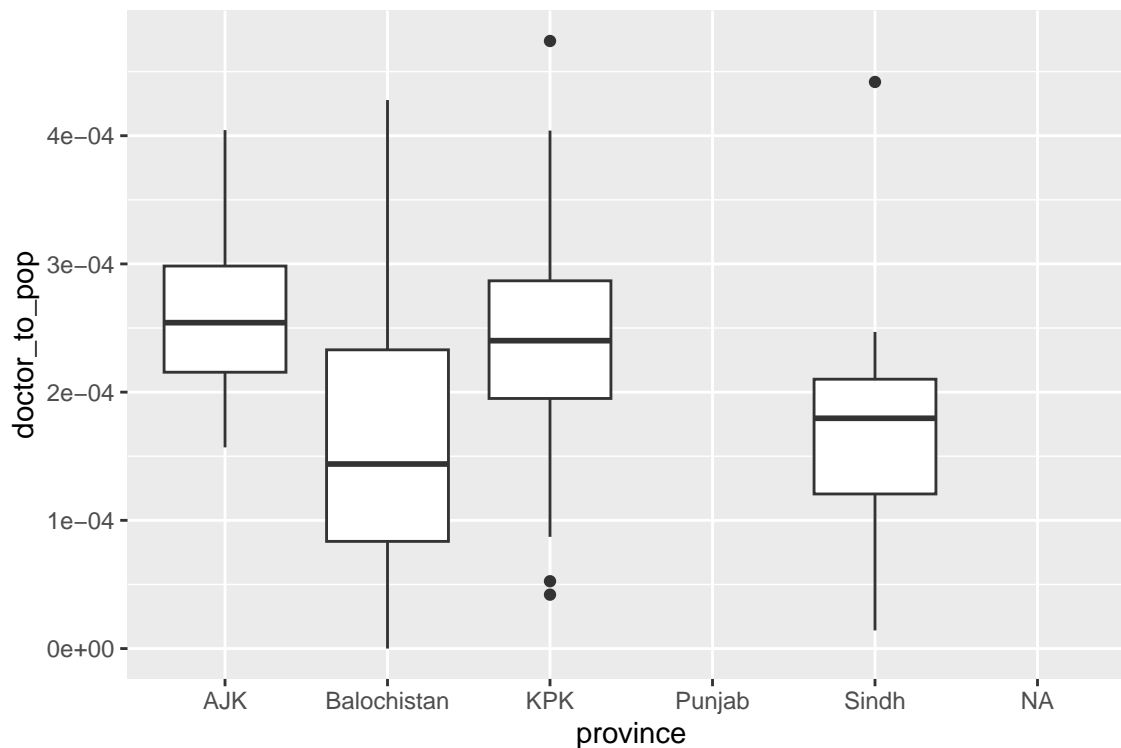
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 37 rows containing non-finite outside the scale range
(`stat_bin()`).



```
districts_two |>
  mutate(
    doctor_to_pop = doctors/pop
  ) |>
  ggplot(aes(x = province, y = doctor_to_pop)) +
  geom_boxplot()
```

Warning: Removed 37 rows containing non-finite outside the scale range (``stat_boxplot()``).



```
model <- glmmTMB(
  doctors ~ poverty + offset(log(pop)) + (1 | province),
  data = districts,
  family = nbinom2
)

tidy(model, conf.int = TRUE) |>
  kable(digits = 3)
```

effect	component	group	term	estimate	std.error	statistic	p.value	conf.low	conf.high
fixed	cond	NA	(Intercept)	-8.064	0.179	-	0.000	-8.414	-7.714
fixed	cond	NA	poverty	-1.333	0.469	-2.843	0.004	-2.252	-0.414
ran_pars	cond	province	sd__(Intercept)	0.067	NA	NA	NA	2.968	5.731

- To go a bit extra, it would be great to talk a bit about multi-level models as this seems like a great type of application that could be used for these. You might think about how to integrate a few of these questions if you have time.

Here are some ideas that might be interesting:

Some initial questions for multi-level models: What are the primary variables considered? What are the level one and level two observational units? What variables are measured at each level?

Some EDA suggestions, some of which you have already implemented in your assignment: 1. Explore general association between the predictor and response variable 2. Explore whether subjects at a given level of the predictor tend to have similar mean responses 3. Explore whether variation in response differs at different levels of a predictor There are a two main ways we can visualize these associations: - A plot of the response versus the predictor for individual observations (meaning treat the observations as independent). - Separate plots of responses versus predictors for each of the level two observations, which is known as a lattice plot.

Based upon these, can you come up with a proposed composite model and write this out. You may want to guide the student through this and help them with it as we have seen it's a bit challenging. Focus on some simplicity and a basic approach.

References

- (1) <https://opendata.com.pk/organization/about/sdpi>
- (2) <https://www.kaggle.com/datasets/alikhan83/pakistan-district-profile>
- (3) <https://opendata.com.pk/dataset/district-profiles-all-districts-of-pakistan>