Quiz 4

```
library(sf)
library(tidyverse)
library(MASS)
library(broom)
library(grid)
```

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

```
localities <- localities |>
  mutate(doctors = `Number of Doctors`,
         poverty = as.numeric(`Multi Dimensional Poverty Index`),
         labor = `Labour force`,
         name = `name`,
         province = Province)


localities <- dplyr::select(localities, doctors, poverty,
                            labor, name, province)
```

Introduction to the Data.

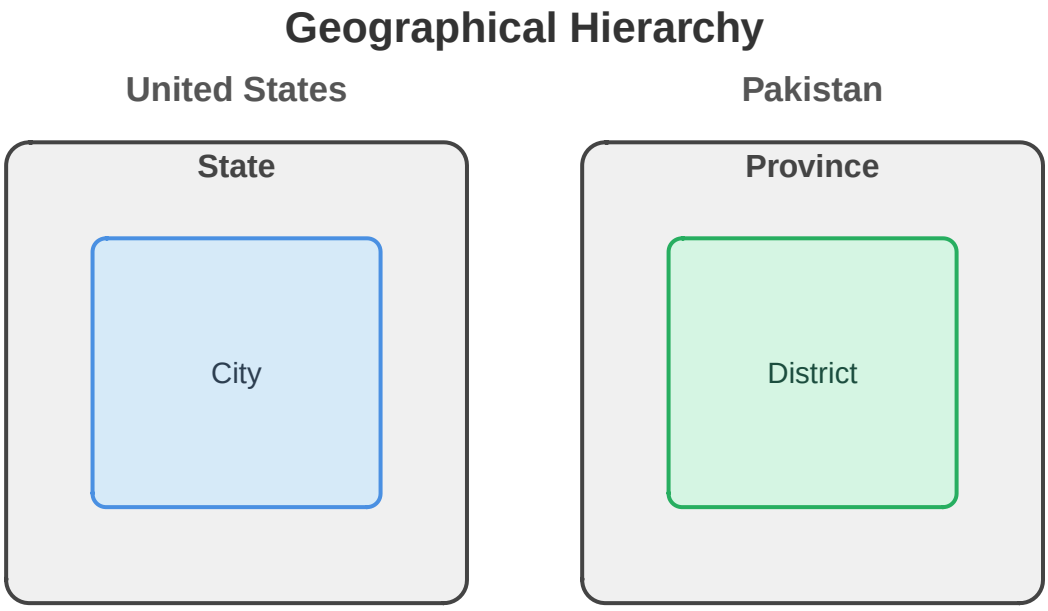
Your task is to model the number of doctors in Pakistan based on poverty levels.

- **localities:** Information about different geographic districts, i.e. cities, 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.
  - *labor*, the size of the labor 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.

```
library(grid)

grid.newpage()

# === Main title ===
grid.text("Geographical Hierarchy", x = 0.5, y = 0.95,
          gp = gpar(fontsize = 16, fontface = "bold", col = "#333333"))

# === Section Labels ===
grid.text("United States", x = 0.25, y = 0.87,
          gp = gpar(fontsize = 13, fontface = "bold", col = "#555555"))
grid.text("Pakistan", x = 0.75, y = 0.87,
          gp = gpar(fontsize = 13, fontface = "bold", col = "#555555"))

# === First nested box (left side) ===

# Outer box 1
grid.roundrect(x = 0.25, y = 0.5, width = 0.4, height = 0.6,
               gp = gpar(col = "#444444", fill = "#F0F0F0", lwd = 2),
               r = unit(0.05, "snpc"))

# Label for outer box 1
grid.text("State", x = 0.25, y = 0.77,
          gp = gpar(fontsize = 12, fontface = "bold", col = "#444444"))

# Inner box 1
grid.roundrect(x = 0.25, y = 0.5, width = 0.25, height = 0.35,
               gp = gpar(col = "#4A90E2", fill = "#D6EAF8", lwd = 2),
               r = unit(0.05, "snpc"))

# Label for inner box 1
grid.text("City", x = 0.25, y = 0.5,
          gp = gpar(fontsize = 11, col = "#2C3E50"))

# === Second nested box (right side) ===

# Outer box 2
grid.roundrect(x = 0.75, y = 0.5, width = 0.4, height = 0.6,
               gp = gpar(col = "#444444", fill = "#F0F0F0", lwd = 2),
               r = unit(0.05, "snpc"))
```

```
# Label for outer box 2
grid.text("Province", x = 0.75, y = 0.77,
          gp = gpar(fontsize = 12, fontface = "bold", col = "#444444"))

# Inner box 2
grid.roundrect(x = 0.75, y = 0.5, width = 0.25, height = 0.35,
               gp = gpar(col = "#27AE60", fill = "#D5F5E3", lwd = 2),
               r = unit(0.05, "snpc"))

# Label for inner box 2
grid.text("District", x = 0.75, y = 0.5,
          gp = gpar(fontsize = 11, col = "#1B4D3E"))
```

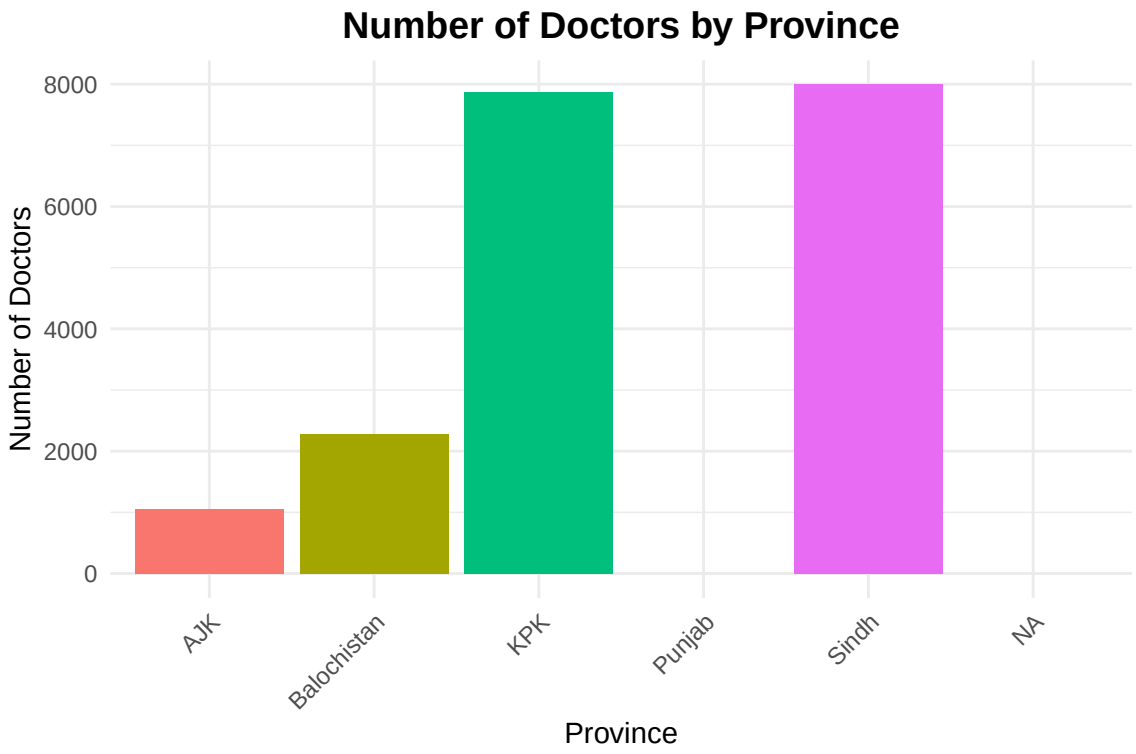


determined how to do grid nested boxes from this conversation: https://chatgpt.com/c/6807c215-16e8-8008-a4ee-e24c79226624

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 n creating a map visual.

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

```r
ggplot(localities, 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:

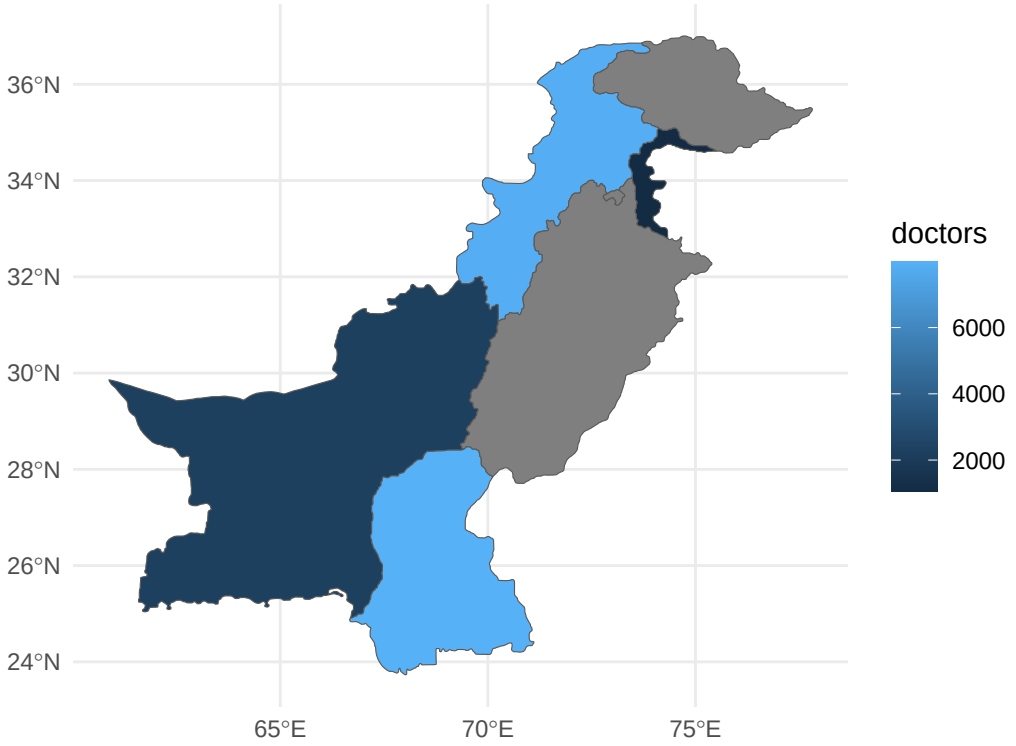
```r
pakistan <- st_read("data/geoBoundaries-PAK-ADM1.shp", quiet = TRUE)
```

```r
provinces <- localities |>
  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()
```



Things to notice:

- Missing data: We don't have information for number of doctors in a couple of provinces.
- Number of 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.

> Additional Instructions.
>
> Before proceeding, we will remove observations where data is missing for doctors, labor, 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.

```
localities <- localities |>
  filter(!is.na(doctors)) |>
  filter(!is.na(labor)) |>
  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(localities$doctors)
```

```
[1] 286.92
```

```
var(localities$doctors)
```

```
[1] 19919.58
```

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 labor 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 <- localities |>
  filter(labor != 0)
```

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

tidy(model, conf.int = TRUE)
```

```
# A tibble: 2 x 7
  term        estimate std.error statistic   p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)    -5.64     0.216    -26.2  7.88e-151    -6.12    -5.15
2 poverty        -1.54     0.728     -2.12 3.43e-  2    -3.17     0.136
```

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 labor 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 labor 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] 0.11
```

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

```
[1] 0.581
```

References

(1) https://www.kaggle.com/datasets/alikhan83/pakistan-district-profile

(2) https://opendata.com.pk/dataset/district-profiles-all-districts-of-pakistan