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 Open Now Pakistan, 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))
```

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

Introduction to the Data.

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

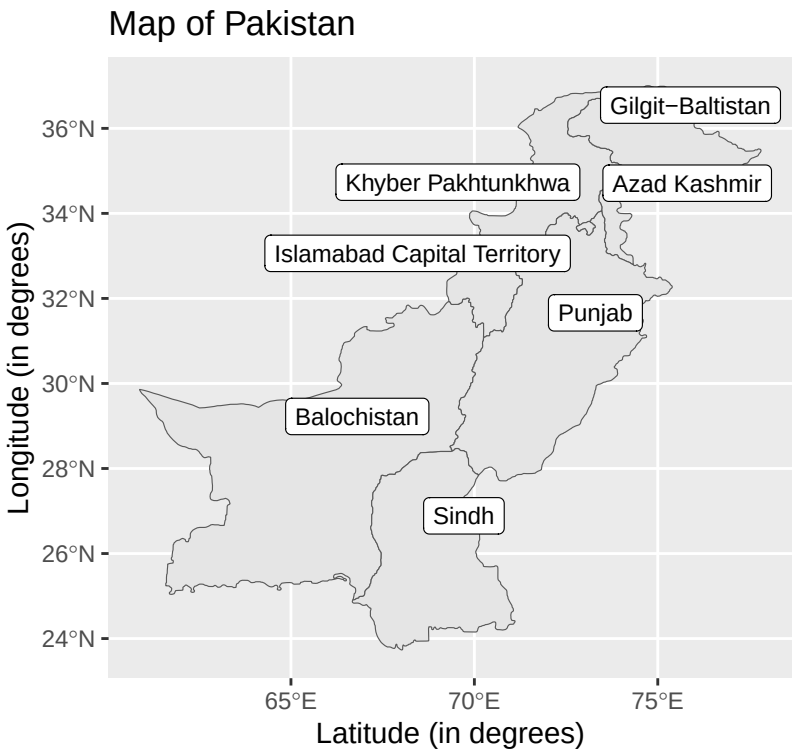
- **districts:** Information about different geographic districts of Pakistan. Dataset was taken from Kaggle (1), which extracted the data from Open Data Pakistan (2).
  - *doctors*, the number of doctors in a district. Our response variable.
  - *pop*, population count
  - *poverty*, a poverty index
  - *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.

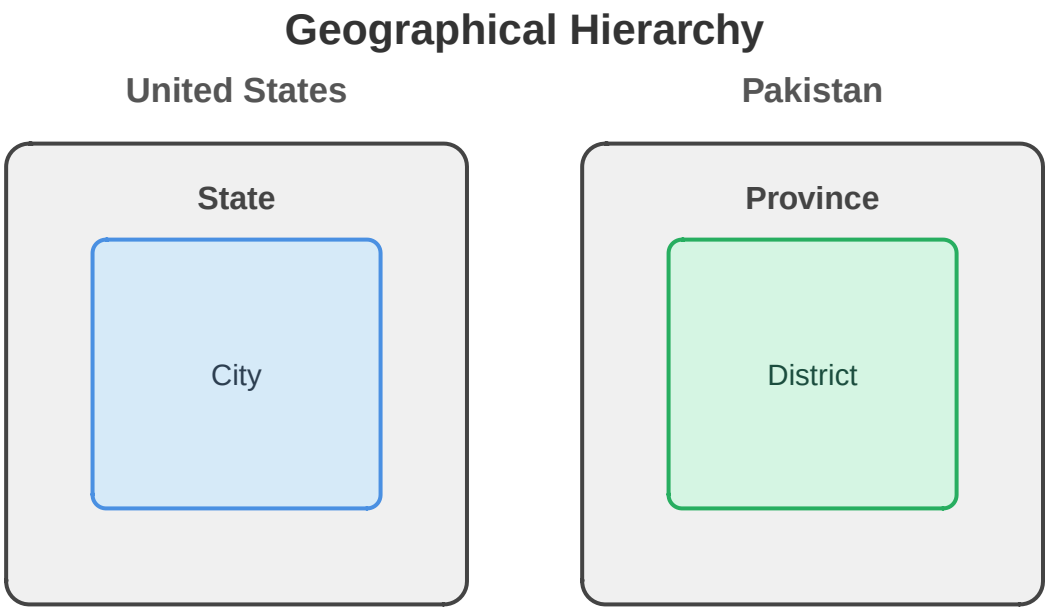
What is a district and province?

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.

```r
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",
    size = 3,
    box.padding = 0.3,
    label.size = 0.2,
    label.r = unit(0.15, "lines"),
    fill = "white",
    color = "black"
  ) +
  labs(
```

```
    title = "Map of Pakistan",
    x = "Latitude (in degrees)",
    y = "Longitude (in degrees)"
  )
```



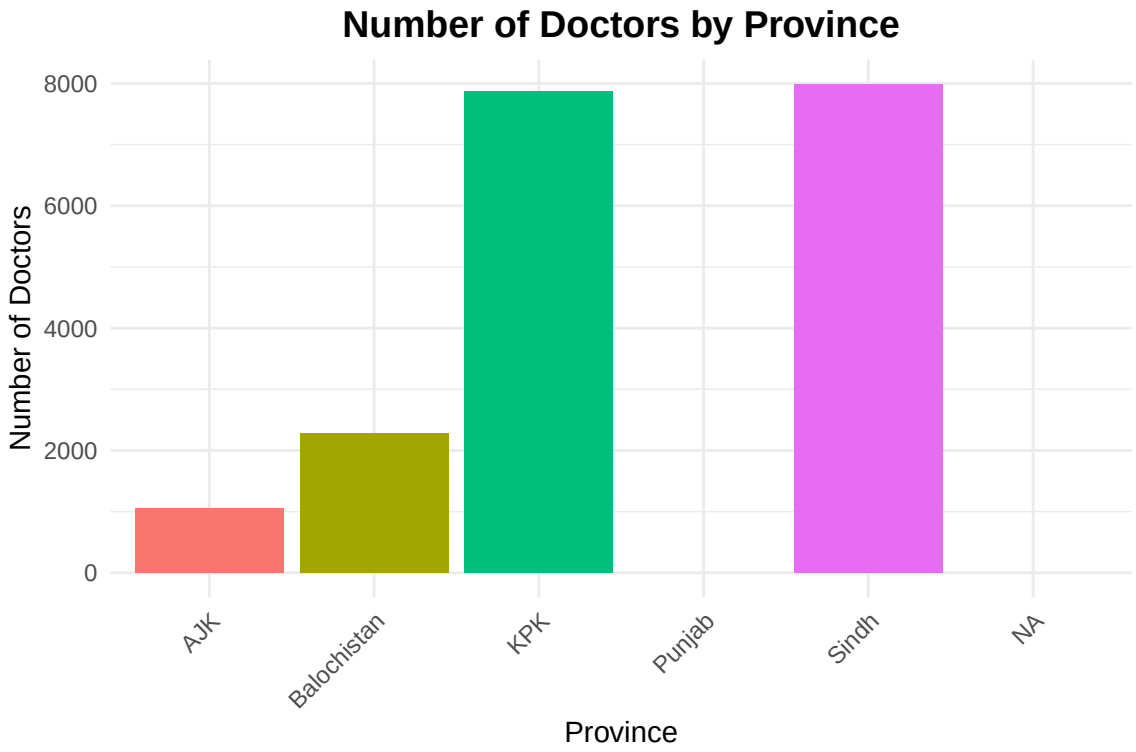

Geographical Hierarchy
United States
State
City
Pakistan
Province
District

Questions: Student Turn

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:

```
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)
  )
```



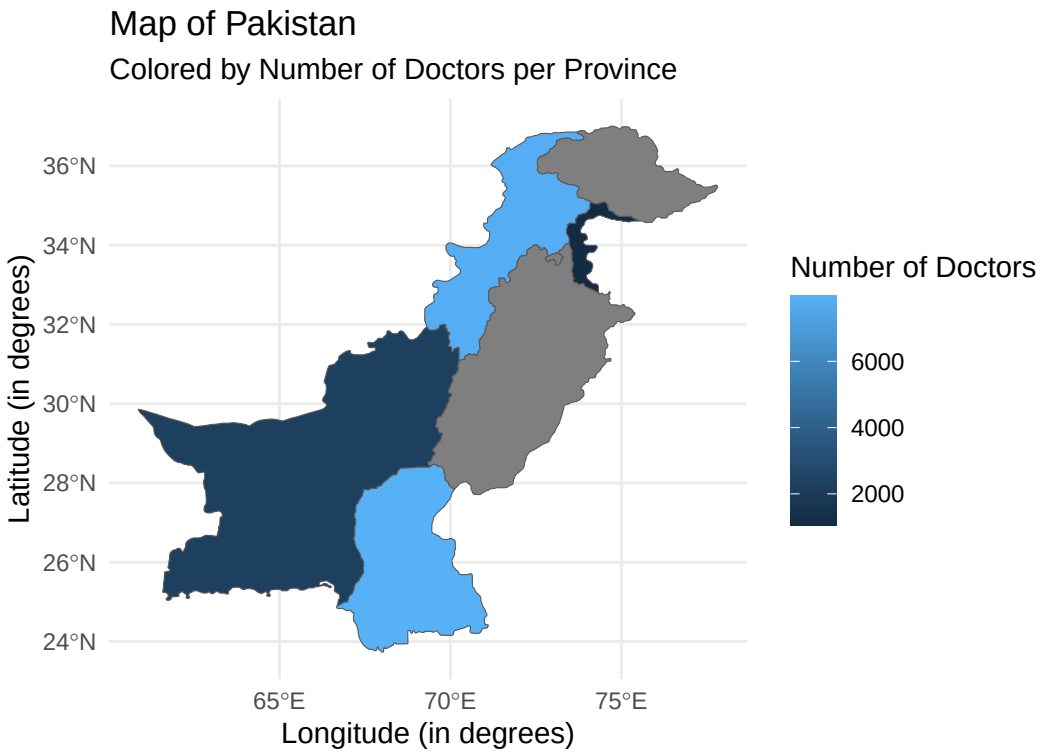
Alternatively, here's another visualization students can make:

```
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. Punjab is dark gray 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.

> 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 population 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 person in a population in a district will change by a multiplicative factor of 0.2143811. This means we expect the number of doctors per person in a population 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 person in a 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)
```

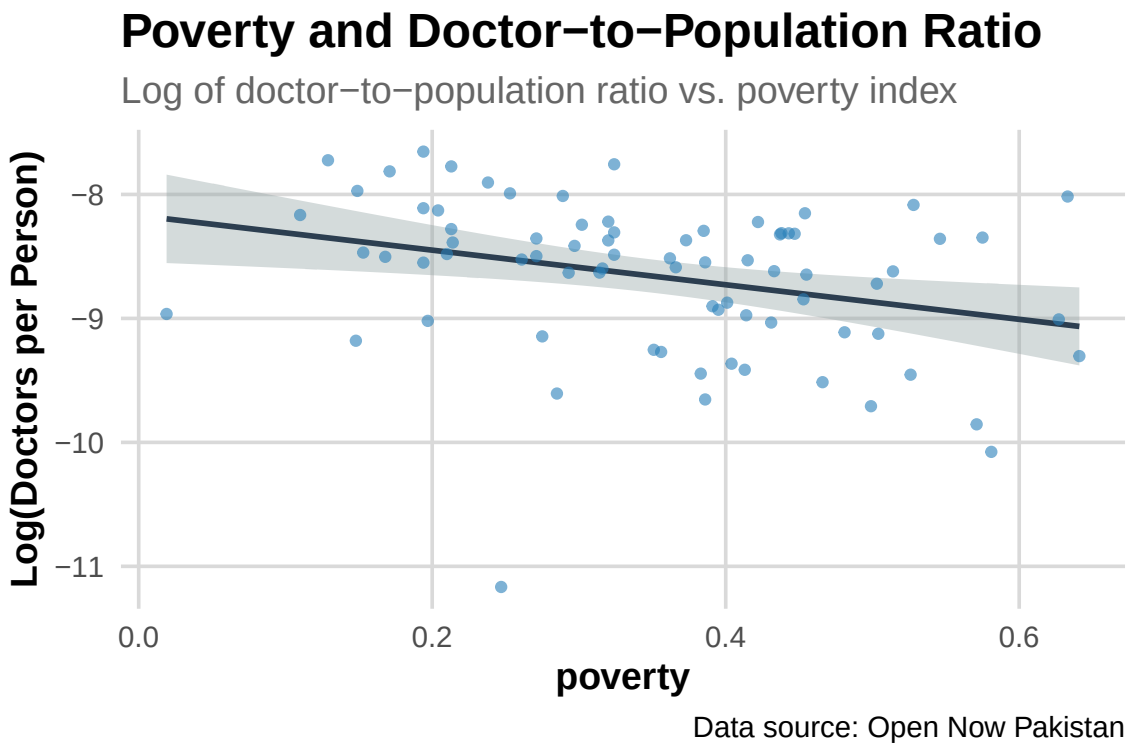
```
[1] 0.641
```

Question 7: Are there obvious level one and level two observational units?

The level one observational unit is a district and the level two observational unit is the province.

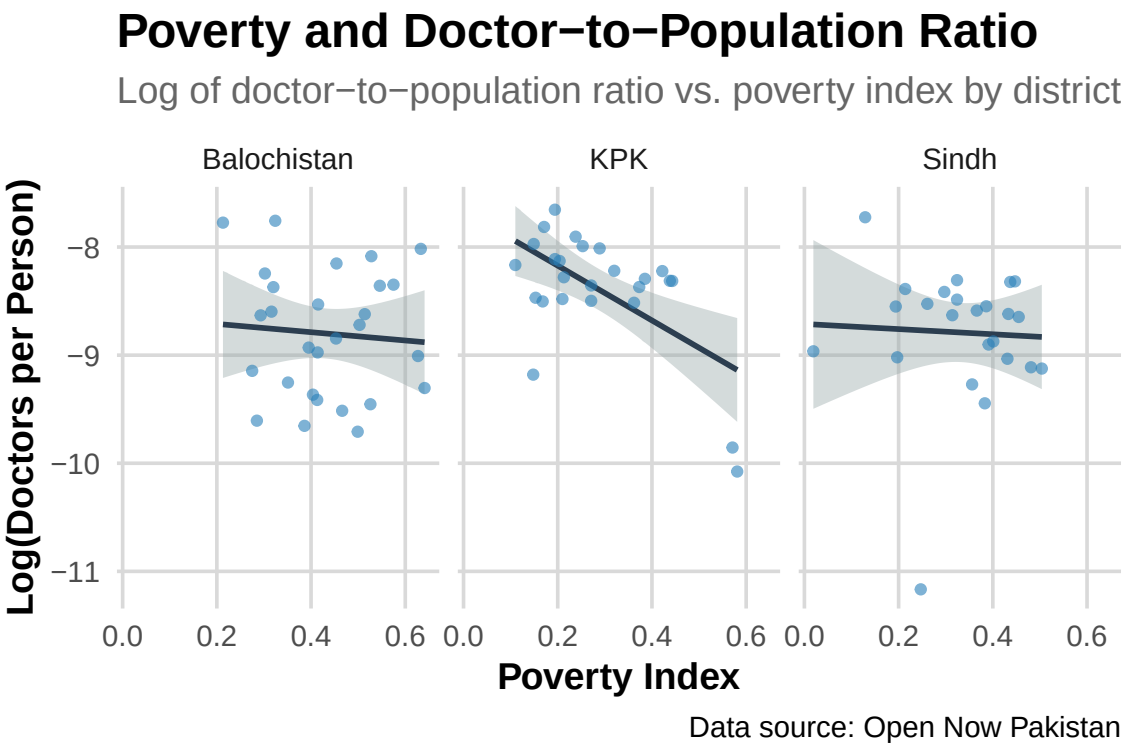
Question 8: Create a visualization to explore the following statement: Districts in the same province in Pakistan have more similar doctor-to-population ratios than those in different provinces. Specifically, plot poverty vs the logarithm of doctors-to-population ratio for all observations. Then, make a visualization showing separate plots for each province.

```
districts |>
  mutate(
    doctor_to_pop = log(doctors / pop)
  ) |>
  ggplot(aes(x = poverty, y = doctor_to_pop)) +
  geom_smooth(method = "lm", se = TRUE, color = "#2C3E50", fill = "#95A5A6") +
  geom_point(alpha = 0.6, color = "#2980B9") +
  labs(
    title = "Poverty and Doctor-to-Population Ratio",
    subtitle = "Log of doctor-to-population ratio vs. poverty index",
    y = "Log(Doctors per Person)",
    caption = "Data source: Open Now Pakistan"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 18),
    plot.subtitle = element_text(size = 14, color = "gray40"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray85"),
    panel.grid.minor = element_blank()
  )
```




```
districts |>
  mutate(
    doctor_to_pop = log(doctors / pop)
  ) |>
  ggplot(aes(x = poverty, y = doctor_to_pop)) +
  geom_smooth(method = "lm", se = TRUE, color = "#2C3E50", fill = "#95A5A6") +
  geom_point(alpha = 0.6, color = "#2980B9") +
  labs(
    title = "Poverty and Doctor-to-Population Ratio",
    subtitle = "Log of doctor-to-population ratio vs. poverty index by district",
    x = "Poverty Index",
    y = "Log(Doctors per Person)",
    caption = "Data source: Open Now Pakistan"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 18),
    plot.subtitle = element_text(size = 14, color = "gray40"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray85"),
```

```
  panel.grid.minor = element_blank()
)+
facet_wrap(~province)
```



Question 9: Fit the multilevel version of the same model that takes into account the variation that occurs across provinces.

```
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	-45.105	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

Question 10: Which estimates are useful for interpretation?

The fixed intercept and fixed effect of poverty.

References

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

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

(3) https://opendata.com.pk/organization/about/sdpi

Use of ChatGPT

ChatGPT was utilized in order to (1) beautify created plots i.e., "Here's my ggplot code, can you beautify this plot?," (2) learn about new plots that exist i.e., use of the grid map to make the visualization of district-province geographical structure and the maps using publicly available GEo IDs,(3) and verify interpretations with the offset and Poisson regression.

Conversation can be found here https://chatgpt.com/c/6807c215-16e8-8008-a4ee-e24c79226624.