Sample-3

For your data project,

1. Introduction: Tell me what problem you are working on? Why is this problem interesting and important. State specific research questions your group will work on. Introduce recent research done in area related to your problem. You can pack all this together to motivate us. Do keep it short, to the point, and interesting.

For this project, I'm looking at public health inequalities and disparities from different countries in the world. There are 4 research questions: How does life expectancy vary across the globe? Do countries with a higher income level have higher life expectancy? Are there any differences in access to medical care between low-income and high-income countries? How does the environmental condition of a country relate to its life expectancy?

A study in 2012 from The Lancet found that there were significant disparities in life expectancy, infant mortality, and disease burden across different countries. In study, one of the determinants of these disparities were income inequality and access to healthcare. In the World Health Statistic report of 2020, also had similar findings in their research. Although life expectancy increased over 8% globally (between years 2000-2016), life expectancy still remained significantly influenced by income.

The research questions reflect the findings of past studies that have been conducted. I'm analyzing recent statistics of countries to see if the issues of disparities are still present. I'm also asking questions to further know if income is also affecting other factors of health, like environmental. Using more recent data what trends I can find that are similar or different to past research.

Data: Tell me about the data resource and explain dimensions of the data, variables in the data, and how does this data relate to your research questions.

I decided on using two datasets. The first data is from the World Health statistics report of 2022, it has the most recent data on health and health related indicators for its 193 member states(https://www.who.int/data/gho/publications/world-health-statistics). The second data s from the World Bank national accounts data, and OECD National Accounts data files, it includes the GDP for all countries(https://data.worldbank.org/indicator/NY.GDP.MKTP.CD). I'm using the first dataset from the World Health organization for the recent statistics for the different variables on public health, I used the first dataset to merge with the first to have the income group of these countries as well. I needed two datasets because I'm are interested in studying whether the health disparities countries may have due to income inequalities.

EDA: Little more advanced

Research Questions: How does life expectancy vary across the globe?

=> Visualization: World Map where the color of each country corresponds with its life expectancy.

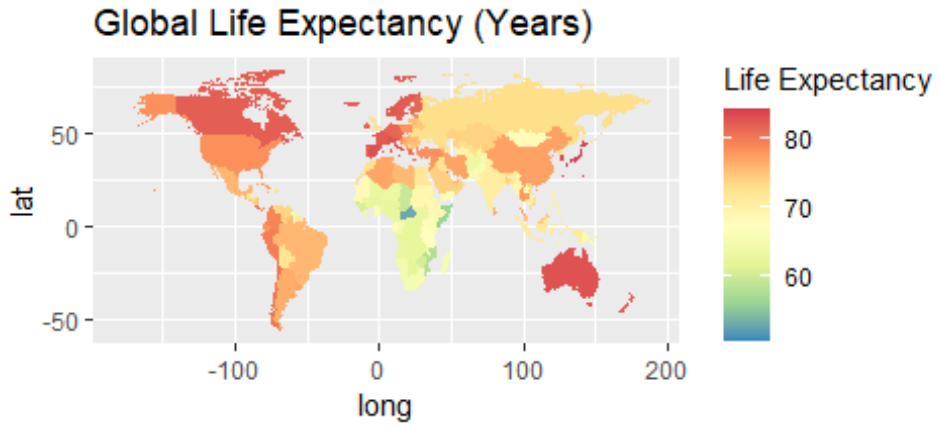
```
world <- map_data("world")

sub = health_gdp_df %>% select(region = 'Country Name', income = IncomeGroup,
life_expectancy = 'Life Expectancy (Box Sexes)')
worldSubset <- inner_join(world, sub, by = "region")
head(worldSubset)
```

```
##       long      lat group order      region subregion     income
## 1 74.89131 37.23164     2    12 Afghanistan      <NA> Low income
## 2 74.84023 37.22505     2    13 Afghanistan      <NA> Low income
## 3 74.76738 37.24917     2    14 Afghanistan      <NA> Low income
## 4 74.73896 37.28564     2    15 Afghanistan      <NA> Low income
## 5 74.72666 37.29072     2    16 Afghanistan      <NA> Low income
## 6 74.66895 37.26670     2    17 Afghanistan      <NA> Low income
##   life_expectancy
## 1            63.2
## 2            63.2
## 3            63.2
## 4            63.2
## 5            63.2
## 6            63.2
```

```
worldle <- ggplot(data = worldSubset, mapping = aes(x = long, y = lat, group
= group)) +
  coord_fixed(1.3) +
  geom_polygon(aes(fill = life_expectancy)) +
  scale_fill_distiller(palette ="Spectral") +
  ggtitle("Global Life Expectancy (Years)")

worldle +labs(fill="Life Expectancy")
```



Research Questions:

Are there any differences in access to medical care between low-income and high-income countries?

Do countries with a higher income level have higher life expectancy?

=> Visualization: Scatterplot of doctor density (x-axis) vs life expectancy (y-axis) colored by income group.

```
#change income group from character to factor variable

health_gdp_df$IncomeGroup = factor(health_gdp_df$IncomeGroup, levels = c("Hig
h income", "Upper middle income", "Lower middle income", "Low income"))

levels(health_gdp_df$IncomeGroup)

## [1] "High income"         "Upper middle income" "Lower middle income"
## [4] "Low income"

g <-ggplot(data=health_gdp_df,aes(x=`Density of medical doctorsw (per 10 000
population)`
                                  ,y=`Life Expectancy (Box Sexes)`)) +geom_po
int(aes(color=IncomeGroup)) +
 labs(x="Density of Medical Doctors (per 10k population)",y="Life Expectancy
(years)",
      title=" Density of Medical Doctors vs Life Expectancy", subtitle="Color
```

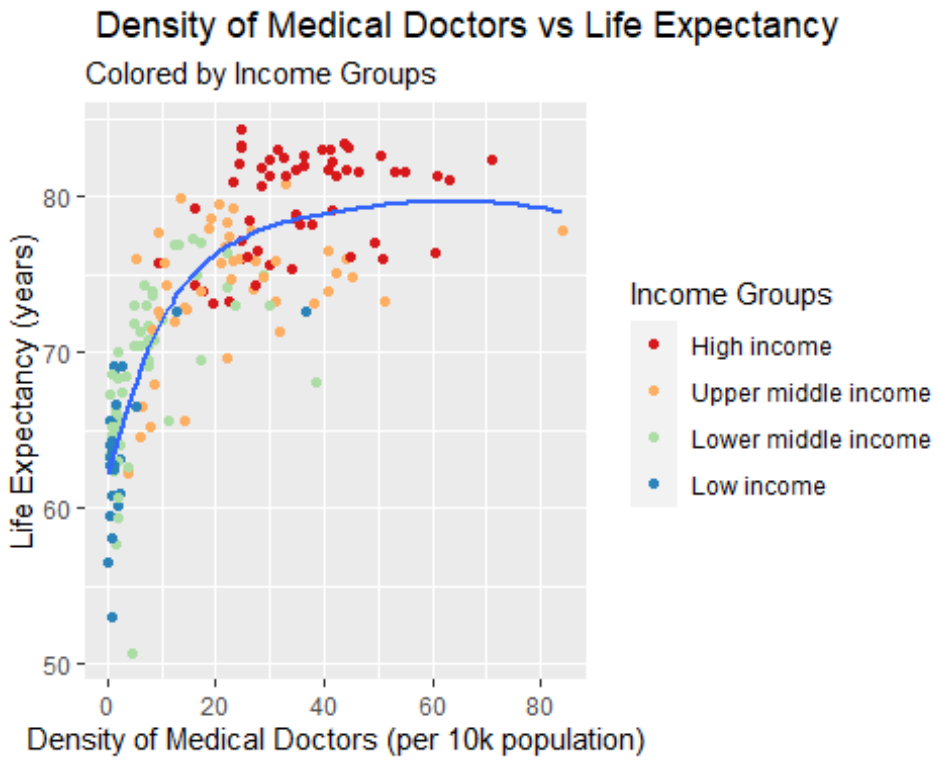
```
ed by Income Groups")  + scale_color_brewer(palette = "Spectral") +labs(color
="Income Groups")

g + geom_smooth(se=F)

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 8 rows containing missing values (`geom_point()`).
```

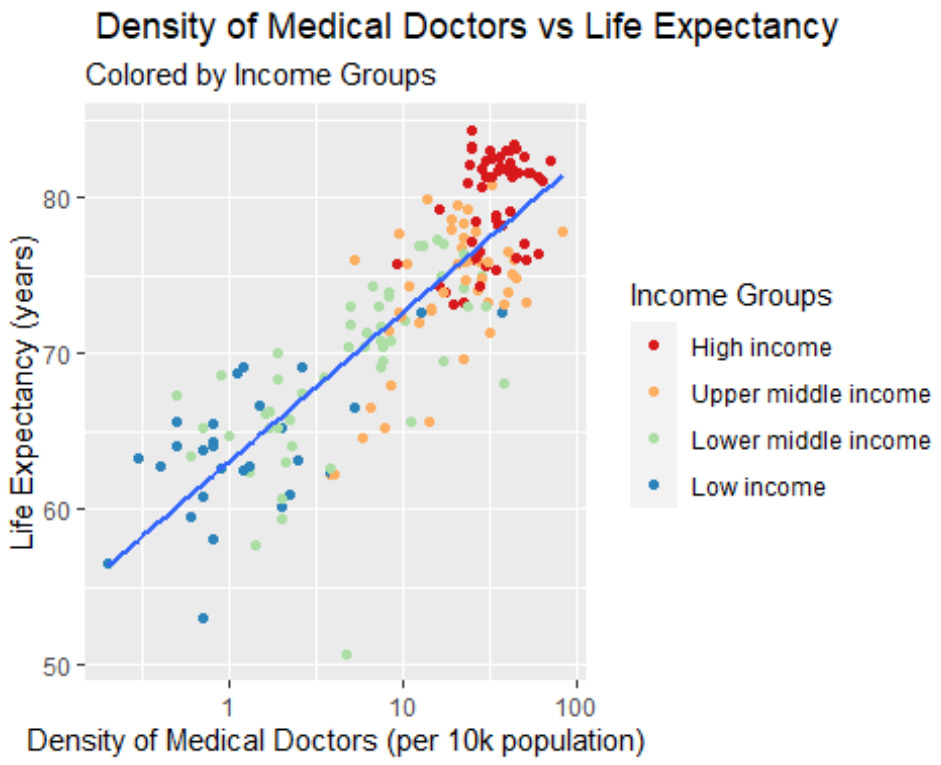



```
#do log transformation on x axis

g_log = g + scale_x_continuous(trans='log10')
g_log +geom_smooth(method = 'lm', se=F)

## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).
## Removed 8 rows containing missing values (`geom_point()`).
```


Density of Medical Doctors vs Life Expectancy
Colored by Income Groups
Life Expectancy (years)
80
70
60
50
1
10
100
Density of Medical Doctors (per 10k population)
Income Groups
High income
Upper middle income
Lower middle income
Low income