Covid19_Report

C. Mayr

2024-05-18

Goal of this Report

In this report I will download data about covid cases in the US. My goal is to figure out if states with higher population density had more cases per person in this state or not.

Loading the data

First step is to include the libraries tidyverse and lubridate. We will need them later in the project.

```r
library(tidyverse)
library(lubridate)
```

Next, load the data from the website, as shown in the lecture. For the analysis we will only need the data sets cases and deaths.

```r
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/" %>%
          str_c("master/csse_covid_19_data/csse_covid_19_time_series/")

filenames <- c("time_series_covid19_confirmed_US.csv",
"time_series_covid19_confirmed_global.csv",
"time_series_covid19_deaths_US.csv",
"time_series_covid19_deaths_global.csv",
"time_series_covid19_recovered_global.csv")

urls <- str_c(url_in, filenames)

#Load relevant US data in Tables
US_cases <- read_csv(urls[1])
US_deaths <- read_csv(urls[3])
```

Prep the data

Here the columns with the dates and cases will be pivoted into rows. We eliminate the columns Lat and Long and convert the date in to the date datatype. At last, the two tables will be fully joined into one.

In the next step the cases and deaths will be summarized by state. This has to happen in two steps. First we take the maximum value of cases, deaths and population and group by county. Afterwards we summarize the numbers for each county and group by state. I save the total population of the USA in the variable total_population. This was a verfication step for myself, after some trouble with the numbers.

To calculate the population density in each state, found a source on github with the areas of each state. I will download this data set, rename the columns, and join the area column to the existing table. Furthermore I mutate the table with the calculated population desity and the cases per person per hundred per state. (This just gives a nicer number to compare)

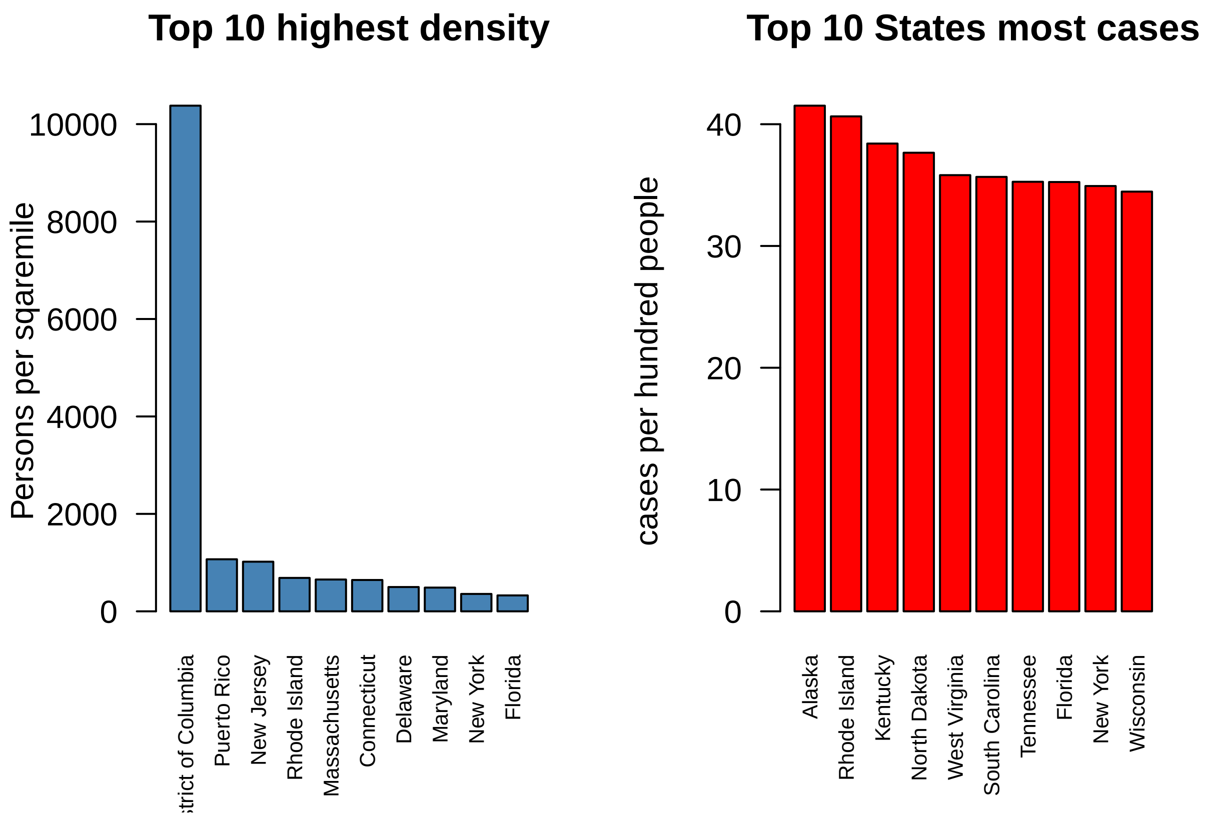
```r
uid_lookup_url <- "https://raw.githubusercontent.com/" %>%
                    str_c("jakevdp/data-USstates/master/state-areas.csv")

AreaData <- read_csv(uid_lookup_url)

AreaData <- AreaData %>%
  rename(Province_State = state,
         area = "area (sq. mi)")

US_Cov_Area <- US_by_State_totals %>%
        full_join(AreaData) %>% filter(area > 0)

US_Cov_Area <- US_Cov_Area %>%
    mutate(PopulationDensity = Population/area,
           CasesPerHundred = (cases*100)/Population)
```

Visualizing Data

To get an first impression on how the data looks and to compare the density data with the area data, I decided to create bar charts for each. I don't want to overload the charts. Therefore, I decided just to display the top ten in each category. If my hypothesis is right, the states with the highest density should also be in the top ten of the most cases. Unfortunately this looks not like if it would be the case.

```r
top_10_states_Dens <- US_Cov_Area %>%
  arrange(desc(PopulationDensity)) %>%
  slice_head(n = 10)

states <- top_10_states_Dens$Province_State
densitys <- top_10_states_Dens$PopulationDensity


par(mfrow = c(1, 2))

barplot(densitys,
        names.arg = states,
        las = 2,  # Rotate the state names for better readability
        col = "steelblue",
        main = "Top 10 highest density",
        xlab = "",
        ylab = "Persons per sqaremile",
        cex.names = 0.7)


top_10_states_cases <- US_Cov_Area %>%
  arrange(desc(CasesPerHundred)) %>%
  slice_head(n = 10)

states <- top_10_states_cases$Province_State
```

```r
cases <- top_10_states_cases$CasesPerHundred

barplot(cases,
        names.arg = states,
        las = 2,  # Rotate the state names for better readability
        col = "red",
        main = "Top 10 States most cases",
        xlab = "",
        ylab = "cases per hundred people",
        cex.names = 0.7)
```

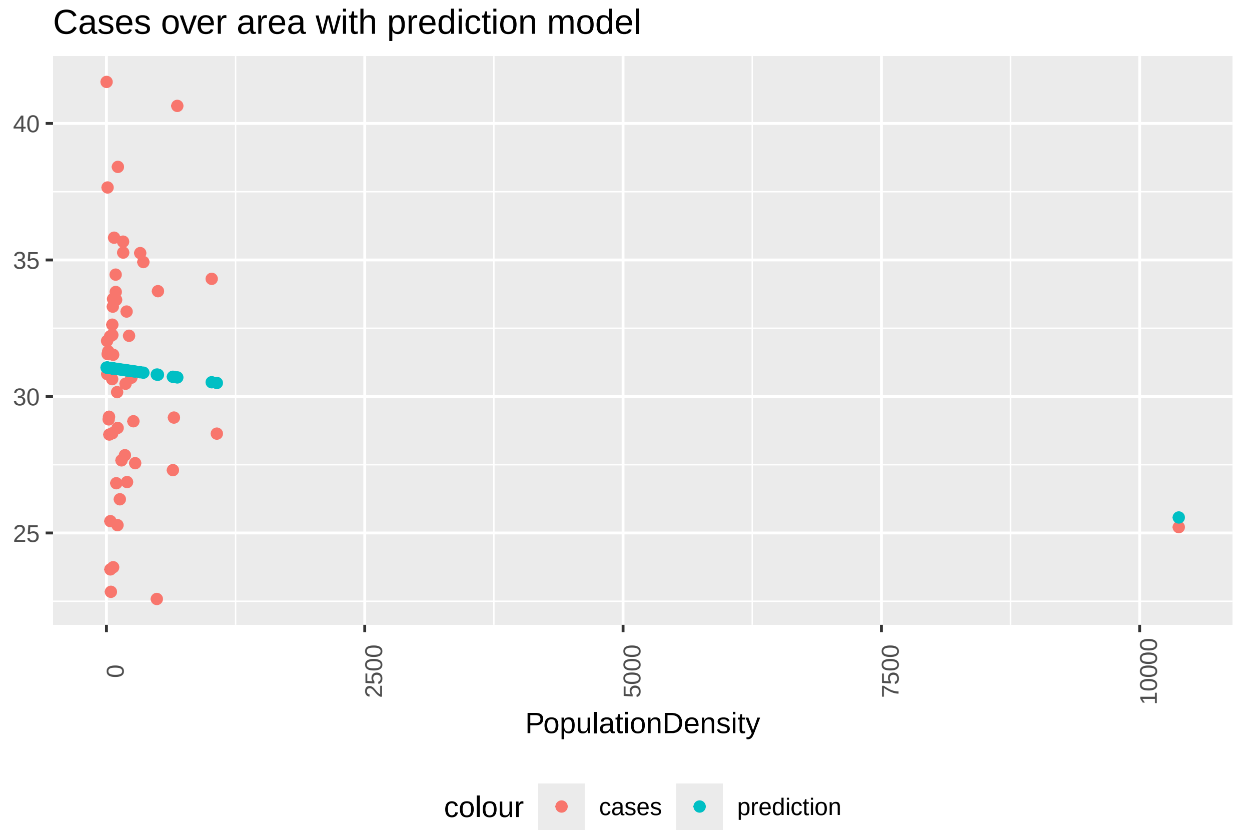


```r
par(mfrow = c(1, 1))
```

Data Model

To get a better overview over all of the states, I implemented a linear model, which predicts the cases over the area.

```r
mod <- lm(CasesPerHundred ~ PopulationDensity, data = US_Cov_Area)

CasesPerSquaremile_withPred <- US_Cov_Area %>%
  mutate(pred = predict(mod))

CasesPerSquaremile_withPred %>%
  ggplot() +
  geom_point(aes(x= PopulationDensity, y= CasesPerHundred, color = "cases")) +
  geom_point(aes(x= PopulationDensity, y= pred, color = "prediction")) +
  theme(legend.position="bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Cases over area with prediction model", y = NULL)
```



Conclusion

By just looking at the data points from the predicted model, it seems like an indirect proportional relationship between the two variables. By looking at the data points one can clearly say there is no correlation between them.

Finally I want to talk about the biases in the data I used to make the analysis.

The data is collected form confirmed covid cases. A lot of people were involved to collect this data and people make mistakes.Furthermore, what is about the unconfirmed cases. For the analysis they would be very interesting and also important to know. I maybe also made a mistake by using the max value and not the increasing rate per day.

The population data can also be biased. In the population data are “only” registered people that live in that area. What about unregistered people. This could also shift the predicted data.