Exploratory Data Analysis

group 25

11/20/2021

Contents

Setup

```
library(tidyverse)
```

Our GitHub Repo: **https://github.com/UBC-MDS/DSCI_522_US_social_determinants_of_health_by_county**

Load Data

```
covid_data <- read.csv("US_counties_COVID19_health_weather_data.csv")
```

Select features

```
interesting_features <- c(
  "date", "county", "cases", "state",
  "total_population", "num_deaths", "percent_smokers",
  "percent_vaccinated", "income_ratio",
  "population_density_per_sqmi", "percent_fair_or_poor_health",
  "percent_unemployed_CHR", "violent_crime_rate",
  "chlamydia_rate", "teen_birth_rate"
)

covid_data <- covid_data %>%
  select(all_of(interesting_features)) %>%
  mutate(date = as.Date(date)) # change date from character to "Date" class
```

```r
# check the descriptive stats of the data frame
summary(covid_data)
```

```
##       date               county              cases           state
##  Min.   :2020-01-21   Length:790331      Min.   :     1   Length:790331
##  1st Qu.:2020-06-01   Class :character   1st Qu.:    29   Class :character
##  Median :2020-08-03   Mode  :character   Median :   174   Mode  :character
##  Mean   :2020-08-02                      Mean   :  1586
##  3rd Qu.:2020-10-04                      3rd Qu.:   768
##  Max.   :2020-12-04                      Max.   :430713
##
##  total_population     num_deaths    percent_smokers  percent_vaccinated
##  Min.   :      76   Min.   :   32   Min.   : 5.909   Min.   : 4.0
##  1st Qu.:   12483   1st Qu.:  235   1st Qu.:14.982   1st Qu.:37.0
##  Median :   27989   Median :  497   Median :17.021   Median :44.0
##  Mean   :  111577   Mean   : 1425   Mean   :17.488   Mean   :42.2
##  3rd Qu.:   75216   3rd Qu.: 1171   3rd Qu.:19.760   3rd Qu.:49.0
##  Max.   :10057155   Max.   :84296   Max.   :41.491   Max.   :66.0
##  NA's   :17835      NA's   :74408   NA's   :17835    NA's   :20649
##   income_ratio    population_density_per_sqmi percent_fair_or_poor_health
##  Min.   : 2.543   Min.   :    0.038           Min.   : 8.121
##  1st Qu.: 4.016   1st Qu.:   19.559           1st Qu.:14.361
##  Median : 4.406   Median :   47.951           Median :17.260
##  Mean   : 4.520   Mean   :  240.895           Mean   :17.953
##  3rd Qu.: 4.874   3rd Qu.:  129.528           3rd Qu.:20.924
##  Max.   :11.971   Max.   :28069.676           Max.   :40.991
##  NA's   :18326    NA's   :17835               NA's   :17835
##  percent_unemployed_CHR violent_crime_rate chlamydia_rate   teen_birth_rate
##  Min.   : 1.302         Min.   :   0.0     Min.   :  35.8   Min.   :  2.11
##  1st Qu.: 3.151         1st Qu.: 121.3     1st Qu.: 230.6   1st Qu.: 18.93
##  Median : 3.885         Median : 209.7     Median : 332.3   Median : 28.15
##  Mean   : 4.135         Mean   : 256.0     Mean   : 404.6   Mean   : 29.71
##  3rd Qu.: 4.815         3rd Qu.: 340.6     3rd Qu.: 505.0   3rd Qu.: 38.97
##  Max.   :19.904         Max.   :1819.5     Max.   :6120.3   Max.   :103.05
##  NA's   :17835          NA's   :61879      NA's   :45401    NA's   :45172
```

Exploratory Data Analysis (EDA)

Table of COVID-19 prevalence for every state

```r
covid_prevalence_table_state <- covid_data %>%

  # The following lines are for calculating daily growth rate
  group_by(state, date) %>%
  summarize(cases = mean(cases),
            population = mean(total_population)) %>%
  mutate(cases_growth_rate = (cases - lag(cases) / lag(cases))) %>%

  # The following lines are for group_by values for each state
  group_by(state) %>%
  summarize(total_cases = max(cases),
            total_cases_per_capita = total_cases / mean(population),
            mean_cases_growth_rate = mean(cases_growth_rate, na.rm=TRUE)) %>%
  arrange(desc(total_cases))
```

```
covid_prevalence_table_state
```

```
## # A tibble: 54 x 4
##    state                total_cases total_cases_per_capita mean_cases_growth_ra~
##    <chr>                      <dbl>                  <dbl>                 <dbl>
##  1 Arizona                   23670.                 0.0244                 7497.
##  2 California                22727.                 0.0205                 6951.
##  3 District of Columbia      22480                  0.0341                10818.
##  4 New Jersey                17033.                 0.0389                 7923.
##  5 Massachusetts             16912.                 0.0322                 6630.
##  6 Connecticut               15863.                 0.0344                 5926.
##  7 Florida                   15467.                 0.0469                 5694.
##  8 Delaware                  12768.                 0.0403                 4728.
##  9 New York                  11899.                 0.0235                 6441.
## 10 Rhode Island              10341                  0.0490                 3648.
## # ... with 44 more rows
```

Table of COVID-19 prevalence for every county

```
covid_prevalence_table_county <- covid_data %>%
  # The following lines are for calculating daily growth rate
  group_by(county, date) %>%
  summarize(cases = mean(cases),
            population = mean(total_population)) %>%
  mutate(cases_growth_rate = (cases - lag(cases) / lag(cases))) %>%

  # The following lines are for group_by values for each state
  group_by(county) %>%
  summarize(total_cases = max(cases),
            total_cases_per_capita = total_cases / mean(population),
            mean_cases_growth_rate = mean(cases_growth_rate, na.rm=TRUE)) %>%
  arrange(desc(total_cases))

covid_prevalence_table_county
```

```
## # A tibble: 1,928 x 4
##    county         total_cases total_cases_per_capita mean_cases_growth_rate
##    <chr>                <dbl>                  <dbl>                  <dbl>
##  1 Los Angeles         430713                 0.0428                140152.
##  2 New York City       329406                 0.0389                200729.
##  3 Miami-Dade          238812                 0.0896                 96144.
##  4 Maricopa            224924                 0.0550                 72659.
##  5 Broward             111629                 0.0599                 43266.
##  6 Cook                107721.                0.0401                 33247.
##  7 Tarrant             107178                 0.0550                 30691.
##  8 San Bernardino      100787                 0.0478                 31933.
##  9 Harris               98320                 0.0414                 36665.
## 10 Riverside            92489                 0.0398                 33880.
## # ... with 1,918 more rows
```

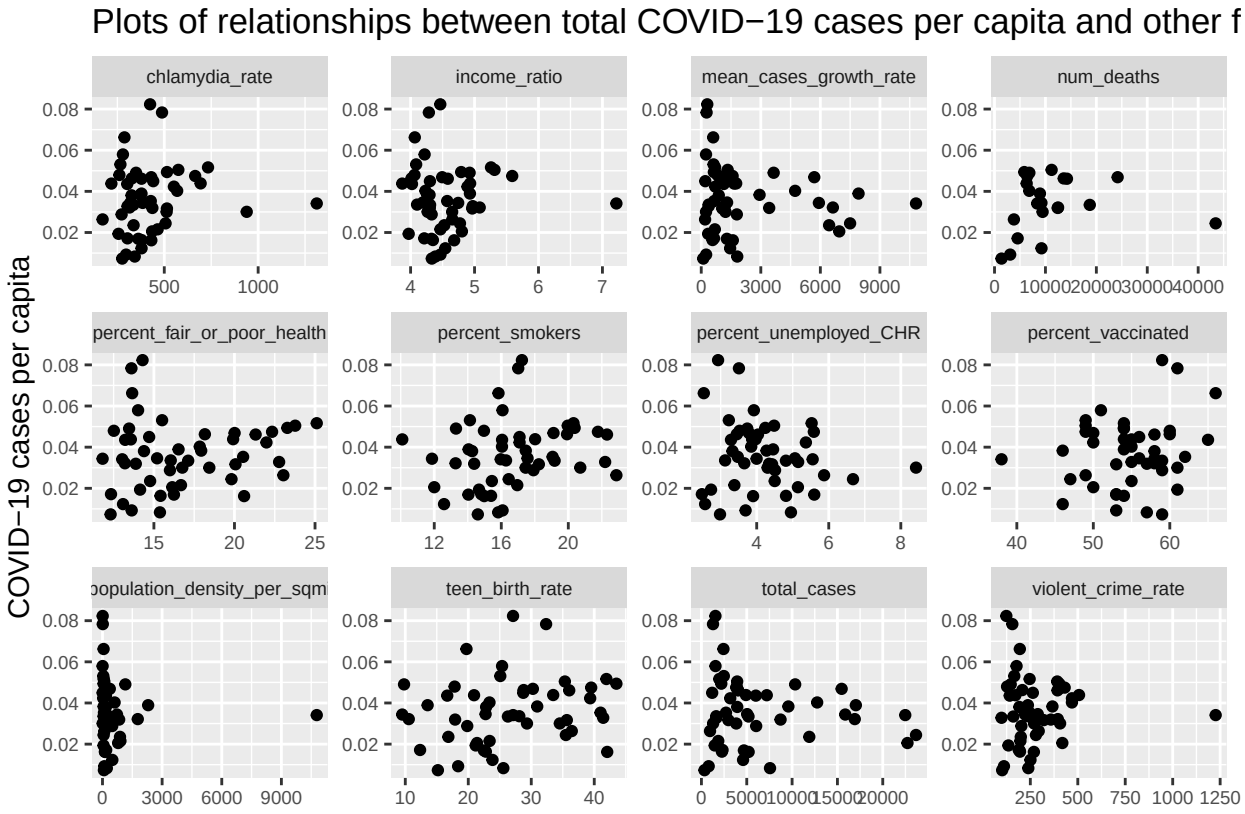
Visualization 1 - relationships between total COVID-19 cases per capita of each state and other features

```
covid_data_group_by_sate <- covid_data %>%
  group_by(state) %>%
  summarize(
            num_deaths = max(num_deaths),
            percent_smokers = mean(percent_smokers, na.rm=TRUE),
            percent_vaccinated = max(percent_vaccinated),
            income_ratio = mean(income_ratio, na.rm=TRUE),
            population_density_per_sqmi = mean(population_density_per_sqmi,
                                               na.rm=TRUE),
            percent_fair_or_poor_health = mean(percent_fair_or_poor_health,
                                               na.rm=TRUE),
            percent_unemployed_CHR = mean(percent_unemployed_CHR, na.rm=TRUE),
            violent_crime_rate = mean(violent_crime_rate, na.rm=TRUE),
            chlamydia_rate = mean(chlamydia_rate, na.rm=TRUE),
            teen_birth_rate = mean(teen_birth_rate, na.rm=TRUE)
            ) %>%
  merge(covid_prevalence_table_state, by="state") %>%
  arrange(desc(total_cases))
```

```
par(mfrow=c(3, 4))

covid_data_group_by_sate_long <- covid_data_group_by_sate %>%
    select_if(is.numeric) %>%
    pivot_longer(-total_cases_per_capita)

covid_data_group_by_sate_long %>%
  ggplot(aes(x=value, y=total_cases_per_capita)) +
  geom_point() +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=7),
        axis.text.y = element_text(size=7)) +
  labs(title="Plots of relationships between total COVID-19 cases per capita and other features",
       x ="",
       y = "COVID-19 cases per capita")
```

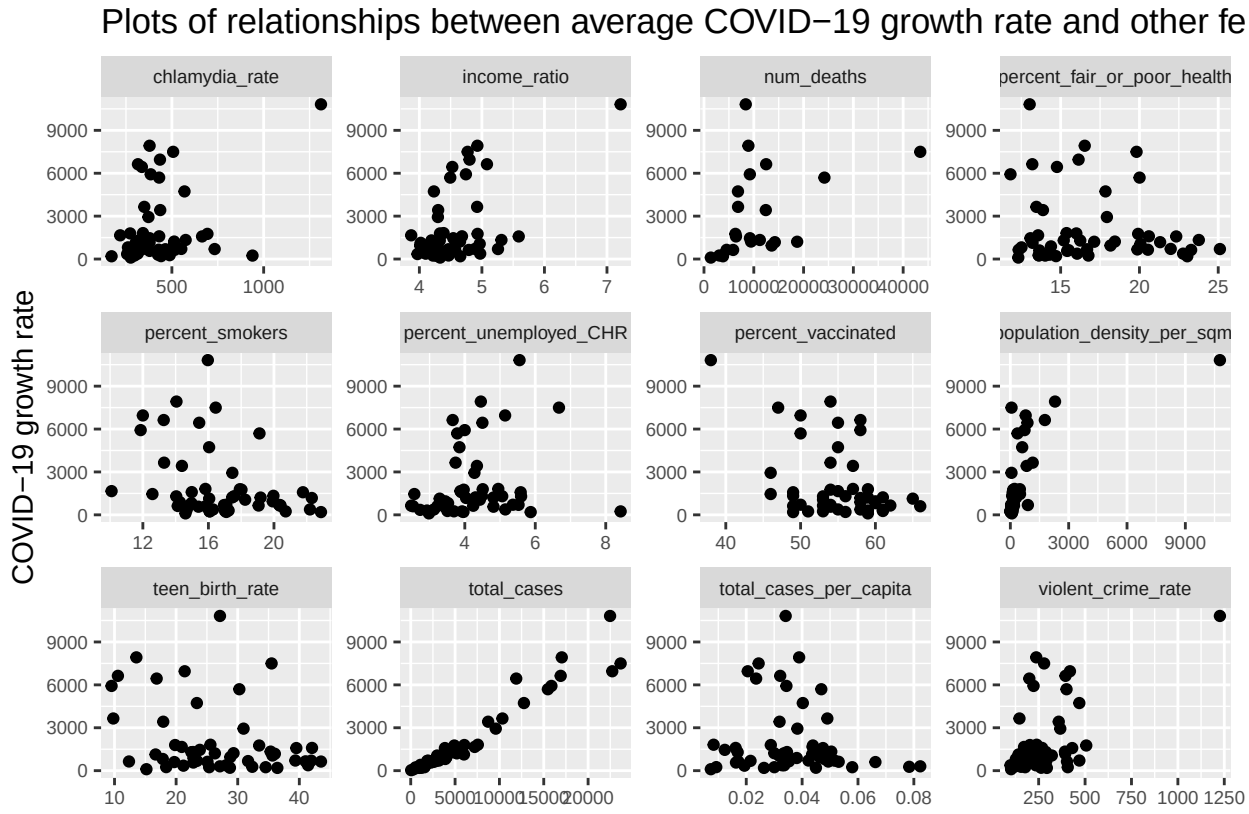


Visualization 2 - relationships between average COVID-19 cases growth rate for each state and other features

```
par(mfrow=c(3, 4))

covid_data_group_by_sate_long <- covid_data_group_by_sate %>%
    select_if(is.numeric) %>%
    pivot_longer(-mean_cases_growth_rate)

covid_data_group_by_sate_long %>%
  ggplot(aes(x=value, y=mean_cases_growth_rate)) +
  geom_point() +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=7),
        axis.text.y = element_text(size=7)) +
  labs(title="Plots of relationships between average COVID-19 growth rate and other features",
       x ="",
       y = "COVID-19 growth rate")
```

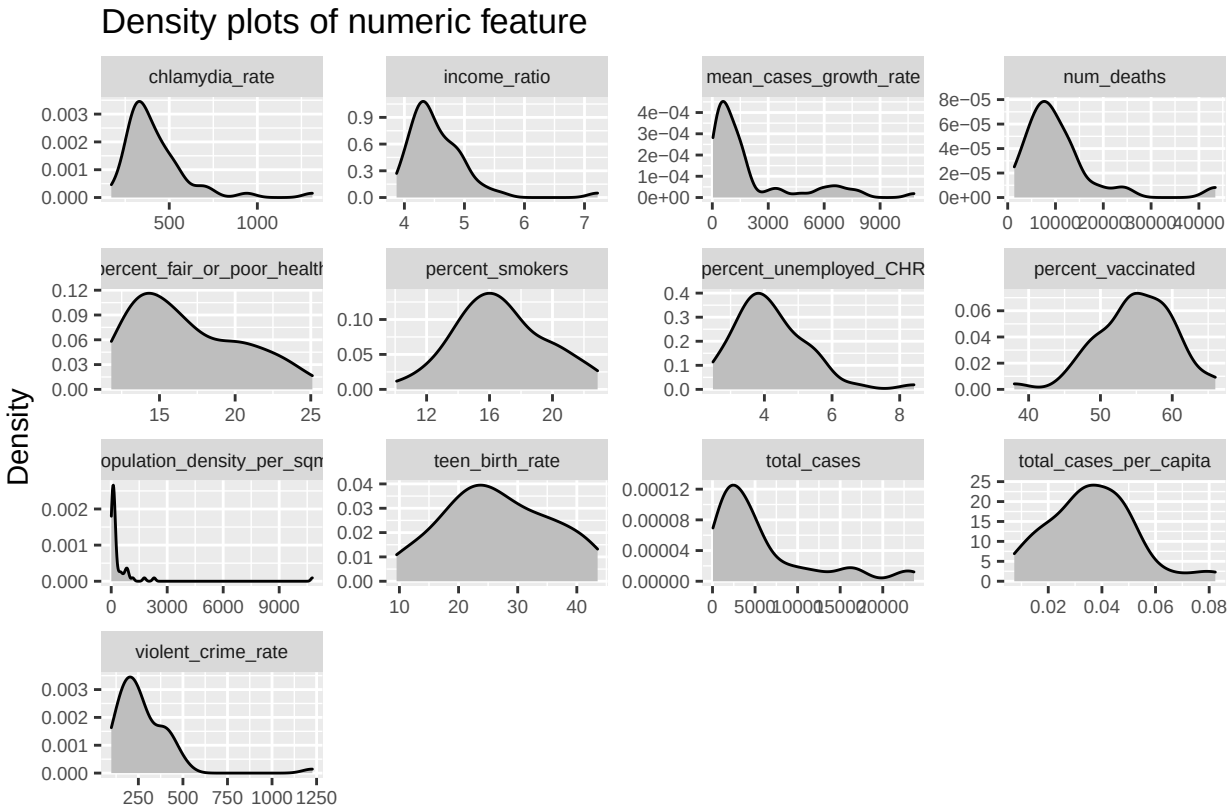


Visualization 3 - distributions of numeric features

```
par(mfrow=c(3, 4))

covid_data_group_by_sate_long <- covid_data_group_by_sate %>%
    select_if(is.numeric) %>%
    pivot_longer(everything())

covid_data_group_by_sate_long %>%
  ggplot(aes(x=value)) +
  geom_density(fill='grey') +
  facet_wrap(~name, scales='free') +
  theme(strip.text = element_text(size=7),
        axis.text.x = element_text(size=7),
        axis.text.y = element_text(size=7)) +
  labs(title="Density plots of numeric feature",
       x ="",
       y = "Density")
```


Density plots of numeric feature
Density
chlamydia_rate
income_ratio
mean_cases_growth_rate
num_deaths
ercent_fair_or_poor_health
percent_smokers
percent_unemployed_CHR
percent_vaccinated
opulation_density_per_sqm
teen_birth_rate
total_cases
total_cases_per_capita
violent_crime_rate