Exploratory Data Analysis

Group 25 - US social determinants of health by county

11/20/2021

Contents

Setup

```
library(tidyverse)
library(plotly)
library(broom)
```

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
```

```
# 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

```
covid_prevalence_table_state <- covid_data %>%

# The following lines are for calculating daily growth rate
  group_by(state, date) %>%
  summarize(cases = sum(cases),
  population = mean(total_population, na.rm=TRUE)) %>%
  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, na.rm=TRUE),
```

```
  mean_cases_growth_rate = mean(cases_growth_rate, na.rm=TRUE)) %>%
  arrange(desc(total_cases))

head(data.frame(covid_prevalence_table_state))
```

```
##        state total_cases total_cases_per_capita mean_cases_growth_rate
## 1      Texas     1322711              4.1147229               399154.6
## 2 California     1318139              1.1911306               401735.0
## 3    Florida     1036294              3.1404685               381475.3
## 4   Illinois      771696              0.8482987               184156.4
## 5   New York      690143              1.3648911               373401.0
## 6    Georgia      473343              5.0353201               168740.8
```

```
tail(data.frame(covid_prevalence_table_state))
```

```
##                       state total_cases total_cases_per_capita
## 49     District of Columbia       22480             0.03411183
## 50                   Hawaii       18373             0.04926750
## 51                    Maine       12833             0.14817337
## 52                  Vermont        4755             0.10257627
## 53           Virgin Islands        1613                    NaN
## 54 Northern Mariana Islands         106                    NaN
##    mean_cases_growth_rate
## 49             10817.9485
## 50              5825.6410
## 51              3959.1199
## 52              1440.5074
## 53               683.7438
## 54                67.8951
```

There are NAs in the table because of the missing values for that county/state in the original dataset.

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 = sum(cases),
  population = mean(total_population, na.rm=TRUE)) %>%
  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, na.rm=TRUE),
  mean_cases_growth_rate = mean(cases_growth_rate, na.rm=TRUE)) %>%
  arrange(desc(total_cases))

head(data.frame(covid_prevalence_table_county))
```

```
##          county total_cases total_cases_per_capita mean_cases_growth_rate
## 1   Los Angeles      430713             0.04282652              140152.12
## 2 New York City      329406             0.03892786              200729.30
## 3          Cook      323162             0.12043222               94207.00
## 4    Miami-Dade      238812             0.08963008               96144.40
```

```
## 5      Maricopa      224924             0.05501316               72659.49
## 6        Harris      196640             0.08270296               73330.53
```

```
tail(data.frame(covid_prevalence_table_county))
```

```
##                          county total_cases total_cases_per_capita
## 1923                    Daggett           9            0.011984021
## 1924                  Petroleum           8            0.017977528
## 1925                     Borden           5            0.007163324
## 1926 Lake and Peninsula Borough           5            0.003543586
## 1927                     Tinian           2                    NaN
## 1928                     Loving           1            0.013157895
##      mean_cases_growth_rate
## 1923              0.7058824
## 1924              3.2833333
## 1925              1.6111111
## 1926              3.4340659
## 1927              0.4615385
## 1928              0.0000000
```

Visualization 1 - distributions of numeric features

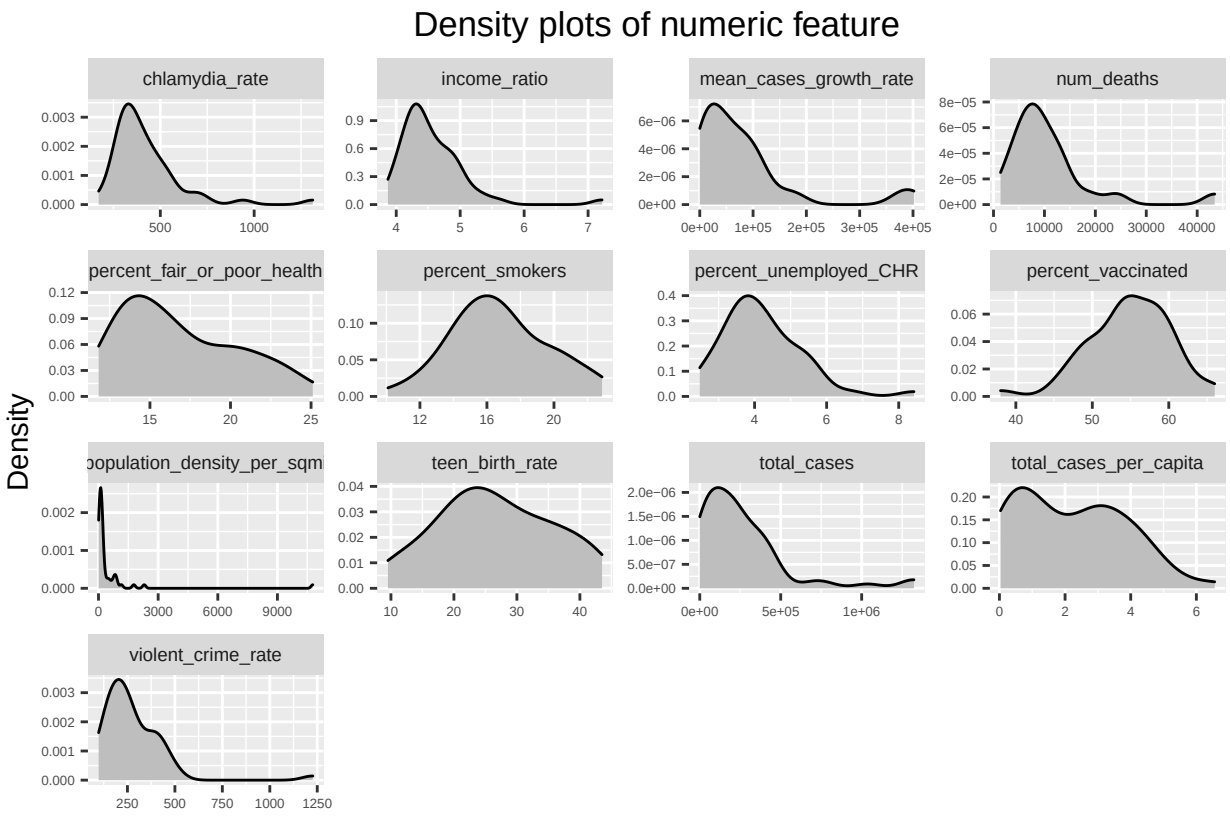
```
covid_data_group_by_sate <- covid_data %>%
  group_by(state) %>%
  summarize(
            num_deaths = max(num_deaths),
            percent_smokers = mean(percent_smokers, na.rm=TRUE),
            percent_vaccinated = max(percent_vaccinated),
            income_ratio = mean(income_ratio, na.rm=TRUE),
            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(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=5),
        axis.text.y = element_text(size=5),
        plot.title = element_text(hjust = 0.5)) +
```

```
  labs(title="Density plots of numeric feature",
       x ="",
       y = "Density")
```



Visualization 2 - relationships between total COVID-19 cases per capita of each state and other features

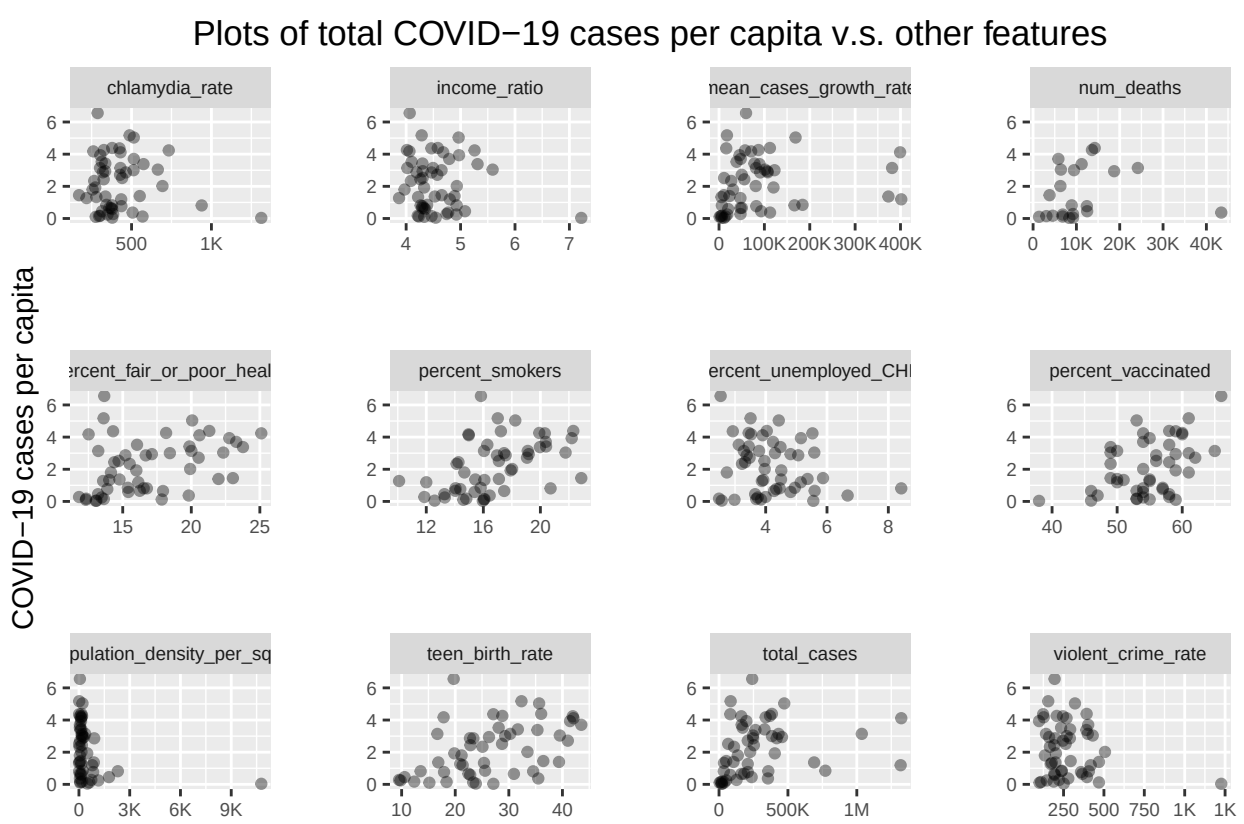
```
par(mfrow=c(3, 4))

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

case_per_capita_plot <- covid_data_group_by_sate_long %>%
  ggplot(aes(x=value, y=total_cases_per_capita)) +
  geom_point(alpha = 0.4) +
  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),
        plot.title = element_text(hjust = 0.5),
        panel.spacing = unit(2.5, "lines")) +
  labs(title="Plots of total COVID-19 cases per capita v.s. other features",
       x ="",
       y = "COVID-19 cases per capita") +
```

```
  scale_x_continuous(labels = scales::label_number_si())


case_per_capita_plot
```



Visualization 3 - 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_growth_rate_plot <- covid_data_group_by_sate_long %>%
  ggplot(aes(x=value, y=mean_cases_growth_rate)) +
  geom_point(alpha = 0.4) +
  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),
        plot.title = element_text(hjust = 0.5),
        panel.spacing = unit(2.5, "lines")) +
  labs(title="Plots of average COVID-19 growth rate v.s. other features",
       x ="",
```

```
       y = "COVID-19 growth rate") +
  scale_x_continuous(labels = scales::label_number_si()) +
  scale_y_continuous(labels = scales::label_number_si())

covid_growth_rate_plot
```