Final Project

Helen Liang

2023-12-20

Introduction

The COVID-19 pandemic has presented unprecedented challenges to public health, economies, and societies worldwide. Understanding the severity and spread of the virus, particularly in hard-hit areas such as the United States and New York, is crucial for policymakers, healthcare professionals, and researchers. It informs strategies for containment, mitigation, and preparedness for future public health emergencies.

To address the problem, I will utilize month-to-month % changes in COVID-19 cases and deaths for the US and New York. Furthermore, the data analysis will involve data Validation to ensure accuracy and consistency in the reported figures, comparative analysis to assess month-over-month changes to highlight trends and spikes, and contextual consideration to acknowledge external factors that can influence the data, such as testing rates and policy changes.

Analytic technique of the analysis will include time-series analysis to understand the dynamics over time and identify patterns and descriptive statistics to summarize the central tendencies, dispersion, and shape of the distributions of percentage changes.

This analysis will provide stakeholders with a clear picture of the pandemic's progression, support evidence-based decisions, and contribute to preparedness for future health crises. The goal is to convert data into insights that will clarify the pandemic's path and consequences.

Packages Required

The following packages are required to be installed and loaded prior to data preparation and data analysis to reproduce the codes and results throughout this project.

```r
## Load Required Packages ##
library(readr) # reading csv data
library(tidyr) # data cleaning
library(dplyr) # data manipulation
library(knitr) # generating tables
library(lubridate) # handling date and time
library(ggplot2) # data visualization
```

Data Preparation

This project involves the utilization of `us.csv` and `us-states.csv` datasets. Prior to assessing how these datasets have behaved over the course of 2020, we must acquire and clean the data.

Loading Data

The `us.csv` and `us-states.csv` datasets are originally sourced from The New York Times GitHub repository: Coronavirus (Covid-19) Data in the United States, where it tracks daily U.S. COVID-19 cases since January

21, 2020, the beginning of the pandemic.

These datasets contain COVID-19 data until January 13, 2021 and are accessible and downloadable through the `GPH-GU-2184 Intermediate Statistical Programming in R` course page on NYU Brightspace.

```
# load us.csv data
us <- read.csv("us.csv")

# load us-states.csv data
states <- read.csv("us-states.csv")
```

The original `us` dataset contains `359` observations and `3` variables. These variables include `date` (daily dates recorded from 01/21/2020 to 01/13/2021), `cases` (cumulative cases), and `deaths` (cumulative deaths). There were no missing (NA's) values in the dataset.

The original `states` dataset contains `17449` observations and `5` variables. These variables include `date` (daily dates recorded from 01/21/2020 to 01/13/2021), `state` (state names), `fips` (standard geographic identifier), `cases` (cumulative cases), and `deaths` (cumulative deaths). Similarly to the `us` dataset, the `states` dataset didn't have any missing (NA's) values.

```
# summary statistics of original "us" data
dim(us)
```

```
## [1] 359   3
```

```
str(us)
```

```
## 'data.frame':    359 obs. of  3 variables:
##  $ date  : chr  "1/21/20" "1/22/20" "1/23/20" "1/24/20" ...
##  $ cases : int  1 1 1 2 3 5 5 5 5 6 ...
##  $ deaths: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
any(is.na(us))
```

```
## [1] FALSE
```

```
# summary statistics of "states" data
dim(states)
```

```
## [1] 17449     5
```

```
str(states)
```

```
## 'data.frame':    17449 obs. of  5 variables:
##  $ date  : chr  "2020-01-21" "2020-01-22" "2020-01-23" "2020-01-24" ...
##  $ state : chr  "Washington" "Washington" "Washington" "Illinois" ...
##  $ fips  : int  53 53 53 17 53 6 17 53 4 6 ...
##  $ cases : int  1 1 1 1 1 1 1 1 1 2 ...
##  $ deaths: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
any(is.na(states))
```

```
## [1] FALSE
```

Data Cleaning

Clean us Data

```
# set "date" column to standardized date format
us$date <- mdy(us$date)
```

```r
# create a new data frame that includes new columns and filters "2020" as the year for data analysis
us1 <- us %>%
  mutate(month = floor_date(date, "month"),
         chr_month = format(date, "%B"),
         year = format(date, "%Y")) %>%
  filter(year == "2020")

head(us1)
```

```
##         date cases deaths      month chr_month year
## 1 2020-01-21     1      0 2020-01-01   January 2020
## 2 2020-01-22     1      0 2020-01-01   January 2020
## 3 2020-01-23     1      0 2020-01-01   January 2020
## 4 2020-01-24     2      0 2020-01-01   January 2020
## 5 2020-01-25     3      0 2020-01-01   January 2020
## 6 2020-01-26     5      0 2020-01-01   January 2020
```

The cleaned and filtered national-level data has been renamed to `us1`. For the `cases` column, the mean is `4992201`, the median is `3289376`, and the standard deviation is `5174886`. For the `deaths` column, the mean is `134142.8`, the median is `134779.5`, and the standard deviation is`99246.17`.

To view the full summary statistics of the data (i.e., mean, median, etc), `summary(us1)` would be the code used to generate the results for all columns in the dataset.

To view each individual value, `mean(us1$cases)` would be used to compute the mean of cases in a R code chunk, `median(us1$cases)` to compute the median of cases, `sd(us1$cases)` to compute the standard deviation of cases, `mean(us1$deaths)` to compute the mean of deaths, `median(us1$deaths)` to compute the median of deaths, and `sd(us1$deaths)` to compute the standard deviation of deaths.

```r
# summary statistics of cleaned and filtered "us1" data
summary(us1)
any(is.na(us1))
```

```r
mean(us1$cases) # mean of cases
```

```
## [1] 4992201
```

```r
median(us1$cases) # median of cases
```

```
## [1] 3289376
```

```r
sd(us1$cases) # sd of cases
```

```
## [1] 5174886
```

```r
mean(us1$deaths) # mean of deaths
```

```
## [1] 134142.8
```

```r
median(us1$deaths) # median of deaths
```

```
## [1] 134779.5
```

```r
sd(us1$deaths) # sd of deaths
```

```
## [1] 99246.17
```

Clean states Data

```
#set "date" column to standardized date format
states$date <- as.Date(states$date, format = "%Y-%m-%d")

# create a new data frame that includes new columns and filters "2020" as the year for data analysis
states1 <- states %>%
  mutate(month = floor_date(date, "month"),
         chr_month = format(date, "%B"),
         year = format(date, "%Y")) %>%
  filter(year == "2020")

head(states1)
```

```
##         date      state fips cases deaths      month chr_month year
## 1 2020-01-21 Washington   53     1      0 2020-01-01   January 2020
## 2 2020-01-22 Washington   53     1      0 2020-01-01   January 2020
## 3 2020-01-23 Washington   53     1      0 2020-01-01   January 2020
## 4 2020-01-24   Illinois   17     1      0 2020-01-01   January 2020
## 5 2020-01-24 Washington   53     1      0 2020-01-01   January 2020
## 6 2020-01-25 California    6     1      0 2020-01-01   January 2020
```

The cleaned and filtered state-level data has been renamed to `states1`. For the `cases` column, the mean is `103221.1`, the median is `29701`, and the standard deviation is `191944.1`. For the `deaths` column, the mean is `2773.6`, the median is `693`, and the standard deviation is`5280.142`. There are no descriptive statistics for the `date`, `month`, `chr_month`, and `year` columns since they either a date or a character column. In addition, descriptive statistics will not be computed for `fips` as it is a geographic identifieer.

To view the full summary statistics of the data (i.e., mean, median, quantiles etc), `summary(us1)` would be the code used to generate the results for all columns in the dataset.

To view each individual value, `mean(states1$cases)` would be used to compute the mean of cases in a R code chunk, `median(states1$cases)` to compute the median of cases, `sd(states1$cases)` to compute the standard deviation of cases, `mean(states1$deaths)` to compute the mean of deaths, `median(states1$deaths)` to compute the median of deaths, and `sd(states1$deaths)` to compute the standard deviation of deaths.

```
# summary statistics of cleaned and filtered "states1" data
summary(states1)
any(is.na(states1))
```

```
mean(states1$cases) # mean of cases
```

```
## [1] 103221.1
```

```
median(states1$cases) # median of cases
```

```
## [1] 29701
```

```
sd(states1$cases) # sd of cases
```

```
## [1] 191944.1
```

```
mean(states1$deaths) # mean of deaths
```

```
## [1] 2773.6
```

```
median(states1$deaths) # median of deaths
```

```
## [1] 693
```

```
sd(states1$deaths) # sd of deaths
```

```
## [1] 5280.142
```

Exploratory Data Analysis

The goal of this project is to investigate the impact of the COVID-19 pandemic on United States and New York by assessing the 2020 monthly percent increase in cases and deaths. We will not be performing any data analysis on 2021 due the fact that we do not have enough information as data is only available until January 13, 2021.

US % Change in COVID-19 Cases Computation

```
us_monthly_cases <- us1 %>%
  group_by(month) %>%
  summarize(begin_cases = first(cases)) %>%
  ungroup()

us_monthly_cases <- us_monthly_cases %>%
  arrange(month) %>%
  mutate(new_cases = lead(begin_cases, default = last(us$cases)) - begin_cases)

us_monthly_cases <- us_monthly_cases %>%
  arrange(month) %>%
  mutate(
    prev_new_cases = lag(new_cases, default = new_cases[1]),
    percent_increase = (new_cases / prev_new_cases) * 100
  ) %>%
  mutate(
    percent_increase = ifelse(is.infinite(percent_increase) | is.nan(percent_increase),
                              NA, percent_increase)
  )
```

New York % Change in COVID-19 Cases Computation

```
states_monthly_cases <- states1 %>%
  group_by(month) %>%
  filter(state == "New York") %>%
  summarize(begin_cases = first(cases)) %>%
  ungroup()

states_monthly_cases <- states_monthly_cases %>%
  arrange(month) %>%
  mutate(new_cases = lead(begin_cases, default = last(states$cases)) - begin_cases)

states_monthly_cases <- states_monthly_cases %>%
  arrange(month) %>%
  mutate(
    prev_new_cases = lag(new_cases, default = new_cases[1]),
    percent_increase = (new_cases / prev_new_cases) * 100
  ) %>%
  mutate(
    percent_increase = ifelse(is.infinite(percent_increase) | is.nan(percent_increase),
                              NA, percent_increase)
  )
```

US and New York % Change in COVID-19 Cases Visualizations

```
us_monthly_cases$month <- as.Date(us_monthly_cases$month)
us_monthly_cases$month_name <- factor(format(us_monthly_cases$month, "%B"),
                                      levels = month.name)

states_monthly_cases$month <- as.Date(states_monthly_cases$month)
states_monthly_cases$month_name <- factor(format(states_monthly_cases$month, "%B"),
                                          levels = month.name)

# table
us_monthly_cases_1 <- us_monthly_cases %>%
  select(month_name, percent_increase)

states_monthly_cases_1 <- states_monthly_cases %>%
  select(month_name, percent_increase)

combined_cases <- merge(us_monthly_cases_1, states_monthly_cases_1, by = "month_name", all = TRUE)

names(combined_cases)[names(combined_cases) == "month_name"] <- "Month"
names(combined_cases)[names(combined_cases) == "percent_increase.x"] <- "US"
names(combined_cases)[names(combined_cases) == "percent_increase.y"] <- "New York"

knitr::kable(combined_cases, caption = "% Change in COVID-19 Cases for US and New York", align = c("c",
```

Table 1: % Change in COVID-19 Cases for US and New York

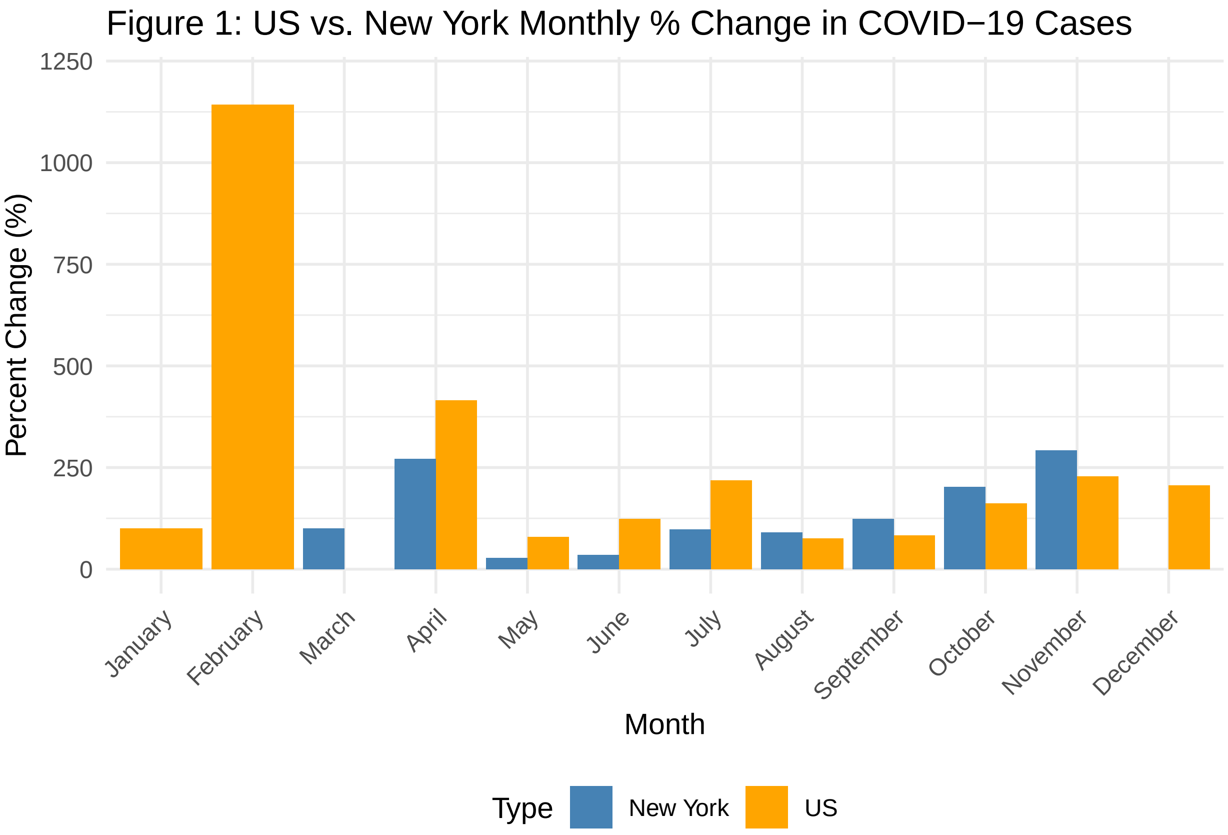
Month	US	New York
January	100.00000	NA
February	1142.85714	NA
March	269128.75000	100.00000
April	415.38437	271.69612
May	79.55330	27.46160
June	123.98180	35.34832
July	218.25275	97.55955
August	75.88854	91.03054
September	83.47996	124.06377
October	161.85293	202.85539
November	228.78139	292.69857
December	206.68808	-420.27769

```
# graph
us_monthly_cases$Type <- 'US'
states_monthly_cases$Type <- 'New York'

combined_data <- rbind(us_monthly_cases, states_monthly_cases)

ggplot(combined_data, aes(x = month_name, y = percent_increase, fill = Type)) +
```

```r
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Figure 1: US vs. New York Monthly % Change in COVID-19 Cases",
       x = "Month", y = "Percent Change (%)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "bottom") +
  scale_fill_manual(values = c("US" = "orange", "New York" = "steelblue")) +
  ylim(0, 1200)
```



US experienced an exceptionally high % increase in cases in the early months of the pandemic, especially noticeable in February and March 2020. However, the % increase remained low from May to October compared to the other months. New York, on the other hand, experienced the sudden % increase in cases once in April and towards the last couple of months, especially in October and November 2020.

The variability in the data could be attributed to various factors, such as public health interventions (i.e., social distancing, the use of face masks, etc.) or changes in data reporting (i.e., over-reporting, under-reporting, etc.).

There are some extreme values (both positive and negative) and missing (NA's) values which will need to be addressed in further analyses to improve on the data presented.

US % Change in COVID-19 Deaths Computation

```r
us_monthly_deaths <- us1 %>%
  group_by(month) %>%
  summarize(begin_deaths = first(deaths)) %>%
  ungroup()

us_monthly_deaths <- us_monthly_deaths %>%
  arrange(month) %>%
  mutate(new_deaths = lead(begin_deaths, default = last(us$deaths)) - begin_deaths)

us_monthly_deaths <- us_monthly_deaths %>%
  arrange(month) %>%
  mutate(
    prev_new_deaths = lag(new_deaths, default = new_deaths[1]),
    percent_increase = (new_deaths / prev_new_deaths) * 100
  ) %>%
  mutate(
    percent_increase = ifelse(is.infinite(percent_increase) | is.nan(percent_increase),
                              NA, percent_increase)
  )
```

New York % Change in COVID-19 Deaths Computation

```r
states_monthly_deaths <- states1 %>%
  group_by(month) %>%
  filter(state == "New York") %>%
  summarize(begin_deaths = first(deaths)) %>%
  ungroup()

states_monthly_deaths <- states_monthly_deaths %>%
  arrange(month) %>%
  mutate(new_deaths = lead(begin_deaths, default = last(states$deaths)) - begin_deaths)

states_monthly_deaths <- states_monthly_deaths %>%
  arrange(month) %>%
  mutate(
    prev_new_deaths = lag(new_deaths, default = new_deaths[1]),
    percent_increase = (new_deaths / prev_new_deaths) * 100
  ) %>%
  mutate(
    percent_increase = ifelse(is.infinite(percent_increase) | is.nan(percent_increase),
                              NA, percent_increase)
  )
```

US vs. New York % Change in COVID-19 Deaths Visualizations

```r
us_monthly_deaths$month <- as.Date(us_monthly_deaths$month)
us_monthly_deaths$month_name <- factor(format(us_monthly_deaths$month, "%B"),
                                       levels = month.name)

states_monthly_deaths$month <- as.Date(states_monthly_deaths$month)
states_monthly_deaths$month_name <- factor(format(states_monthly_deaths$month, "%B"),
                                           levels = month.name)

# table
us_monthly_deaths_1 <- us_monthly_deaths %>%
  select(month_name, percent_increase)

states_monthly_deaths_1 <- states_monthly_deaths %>%
  select(month_name, percent_increase)

combined_deaths <- merge(us_monthly_deaths_1, states_monthly_deaths_1, by = "month_name", all = TRUE)

names(combined_deaths)[names(combined_deaths) == "month_name"] <- "Month"
names(combined_deaths)[names(combined_deaths) == "percent_increase.x"] <- "US"
names(combined_deaths)[names(combined_deaths) == "percent_increase.y"] <- "New York"

knitr::kable(combined_deaths, caption = "% Change in COVID-19 Deaths for US and New York", align = c("c
```

Table 2: % Change in COVID-19 Deaths for US and New York

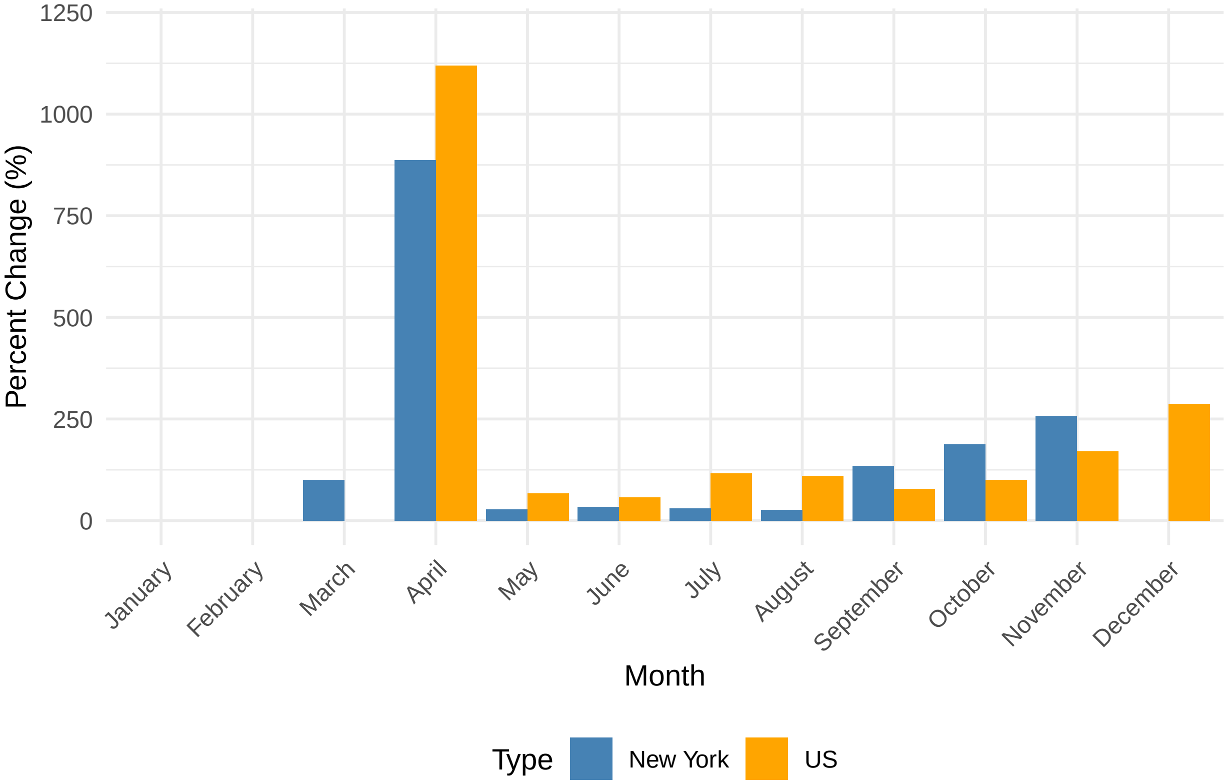
Month	US	New York
January	NA	NA
February	NA	NA
March	177400.00000	100.00000
April	1119.44758	887.20497
May	67.49417	27.65332
June	57.17590	34.17722
July	116.62390	29.58025
August	110.56577	26.87813
September	78.04088	134.78261
October	100.44087	187.09677
November	170.82365	258.12808
December	287.61336	-3215.64885

```r
# graph
us_monthly_deaths$Type <- 'US'
states_monthly_deaths$Type <- 'New York'

combined_data <- rbind(us_monthly_deaths, states_monthly_deaths)

ggplot(combined_data, aes(x = month_name, y = percent_increase, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Figure 2: US vs. New York Monthly % Change in COVID-19 Deaths",
       x = "Month", y = "Percent Change (%)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "bottom") +
  scale_fill_manual(values = c("US" = "orange", "New York" = "steelblue")) +
  ylim(0, 1200)
```

Figure 2: US vs. New York Monthly % Change in COVID−19 Deaths



US experienced an exceptionally high % increase in deaths in the early months of the pandemic, especially noticeable in March and April 2020. However, the % increase remained low from May to October compared to the other months. New York experienced high % increase in deaths in April and November 2020.

The variability in the data could be attributed to various factors, such as rapid improvements in healthcare to address the COVID-19 pandemic (i.e., vaccinations, medications, etc.) or changes in data reporting (i.e., over-reporting, under-reporting, etc.).

There are some extreme values (both positive and negative) and missing (NA's) values which will need to be addressed in further analyses to improve on the data presented.

Summary

The outbreak of the COVID-19 pandemic was detected from the steep % increases in early 2020 for the US and New York, indicative of the virus's rapid spread. The US recorded its peak increases in February and March, while New York's numbers soared in April and fluctuated thereafter. The data illustrated New York's situation improved relative to the rest of the US, where variability in case changes persisted, suggesting the influence of public health measurees and the pandemic's evolution.

However, the analysis is limited by potential inconsistencies in testing and reporting, presented by having missing (NA's) and extreme (both positive and negative) values. It's also lacking other variables such as population density or public health policies. Future research should incorporate these factors as well as extending the timeframe beyond 2020 to have a more comprehensive analysis.