# COVID Project

Xiangnan Chen & Augistine Cui & Yiyuan Wang

2020/9/14

## R Markdown

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

### Introduction

COVID-19 is an infectious disease caused by the most recently discovered coronavirus. The disease spreads primarily from person to person through small droplets from the nose or mouth when people with COVID-19 cough, sneeze or speak [1]. The current coronavirus disease COVID-19 pandemic is hitting the globe unprecedentedly. Lives have been taken, and economic activities have been stagnated. Thus, it is crucial for people to understand better this global pandemic's current situation and future development to be better prepared to solve this global crisis. In this report, we are using COVID-19 data gathered in the United States to make predictions for the future trend of COVID-19 and check if there is a relationship between date and cummulative cases.

### Background

Among all the studies that focus on COVID-19, we find three sources interesting. The first is CDC's own weekly report, it contains various data that can give us overview of the current situation. Another two researches analyze two different aspects. The data from Utah analyze the relationship between COVID-19 cases, hospitalization and Testing with level of deprivation in different area and relations to different races. Last research focuses on the Age distribution of the COVID-19 pandemic.

The CDC provides detailed weekly summary on the dataset in order to keep track and study the tendency of the COVID-19. This report also combines some other data source that CDC collects. Like data of Public Health lab and influenza-like illness (which has similar symptom compatible with COVID-19). The majority part related to the dataset we would like to use are hospitalization. The data have been cleaned with missing/impropriate value with Unknown, then analyze the laboratory confirmed COVID-19-associated hospitalization of each age group each week. The relationship between race and hospitalization are also analyzed. The overall cumulative COVID-19 hospitalization rate is 174.8 per 100,000, with the highest rates in people aged

65 years and older (472.3 per 100,000) and 50–64 years (261.5 per 100,000). Also, Hispanic or Latino have the highest hospitalization rate (358.5 per 100,000)[2].

The analysis on Utah is a regional analyze on COVID-19 Data. It analyzes cases, Hospitalization, Testing in different area of Utah states. Data mostly comes from Utah Department of Health. It analyzes the relationship between cases, Hospitalization and Testing and the level of deprivation. The level of deprivation also tells the majority races within that area. They find that the infection of high-deprivation areas of Utah are three times higher than the lower-deprivation areas, so does the rates of hospitalization and testing[3]. Those area are characterized by large proportion of Hispanic and Latino residents (similar to the CDC report above)

The last research focus on the relationship between COVID-19 and ages. The first source states that highest infection and hospitalization rates are among older adults. The source researches the relationship of incidence in each age group with each month. And the weekly median age of people with COVID-19. It finds that the distribution tends to young adults from 20-39 years, younger adults are likely to contribute to community transmission of COVID-19[4].

### Data

```r
cov <- read.csv("COVID-19_Case_Surveillance_Public_Use_Data.csv")
```

```r
head(cov, 10)
```

```
##    cdc_report_dt pos_spec_dt   onset_dt            current_status     sex
## 1     2020/03/03  2020/03/03            Laboratory-confirmed case    Male
## 2     2020/03/03  2020/03/03            Laboratory-confirmed case  Female
## 3     2020/04/07  2020/03/03 2020/03/03 Laboratory-confirmed case Unknown
## 4     2020/08/04  2020/08/04                        Probable Case    Male
## 5     2020/07/28  2020/08/04 2020/07/28 Laboratory-confirmed case    Male
## 6     2020/08/03  2020/08/04 2020/08/03 Laboratory-confirmed case    Male
## 7     2020/08/04  2020/08/04            Laboratory-confirmed case    Male
## 8     2020/08/04  2020/08/04            Laboratory-confirmed case    Male
## 9     2020/08/04  2020/08/04            Laboratory-confirmed case    Male
## 10    2020/08/04  2020/08/04            Laboratory-confirmed case    Male
##      age_group Race.and.ethnicity..combined. hosp_yn  icu_yn death_yn
## 1  0 - 9 Years                       Unknown Missing Missing  Missing
## 2  0 - 9 Years                       Unknown Missing Missing  Missing
## 3  0 - 9 Years                       Unknown      No Missing  Missing
## 4  0 - 9 Years                       Unknown Missing Missing  Missing
## 5  0 - 9 Years                       Unknown      No      No       No
## 6  0 - 9 Years                       Unknown Missing Missing  Missing
## 7  0 - 9 Years                       Unknown Unknown Unknown       No
## 8  0 - 9 Years                       Unknown Unknown Unknown       No
## 9  0 - 9 Years                       Unknown Unknown Unknown       No
## 10 0 - 9 Years                       Unknown Missing Missing  Missing
##    medcond_yn
## 1     Missing
## 2     Missing
## 3     Missing
## 4     Missing
## 5     Missing
## 6     Missing
## 7     Unknown
## 8     Unknown
## 9     Unknown
## 10    Missing
```

There are three quntitative variables ‘cdc_report_dt’, ‘pos_spec_dt’ and ‘onset_dt’. Thhere are eight categorical varaiables ‘current_status’, ‘sex’, ‘age_group’, ‘Race.and.ethnicity..combined’,‘hosp_yn’, ‘icu_yn’ ,‘death_yn’ and ‘medcond_yn’. Dscription of each variable is in the Appendix.

```
nrow(cov)
```

```
## [1] 3662325
```

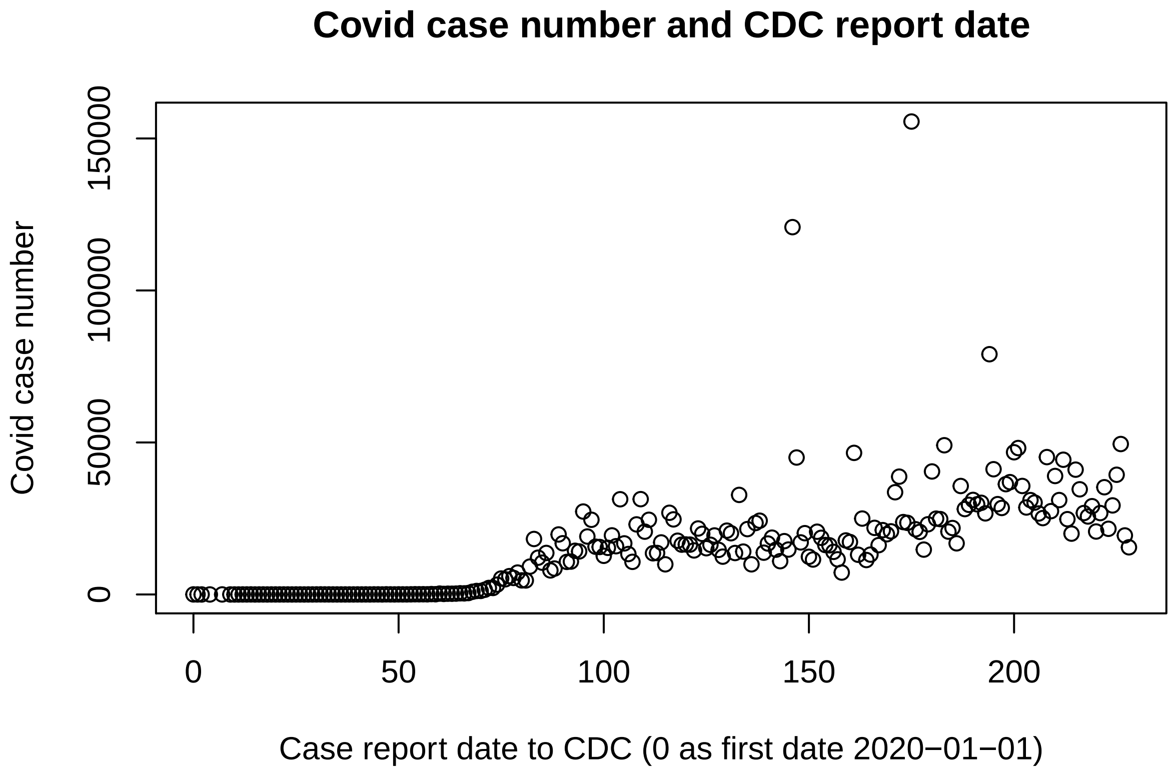
```
ncol(cov)
```

```
## [1] 11
```

There are 3662325 observations and each observation contians 11 variables.

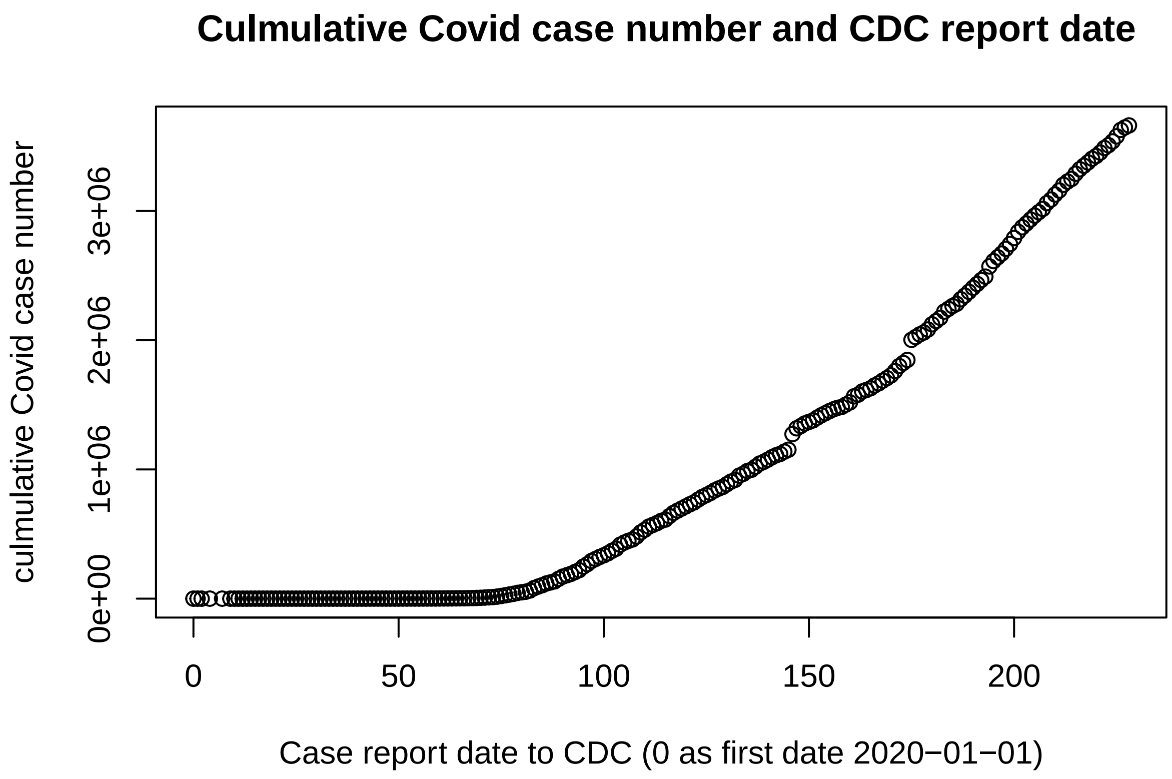
```
reportDate <- unique(cov$cdc_report_dt)
case_date <- count(cov, cov$cdc_report_dt)
case_date$date <- as.Date(case_date$`cov$cdc_report_dt`)
case_date$date <- case_date$date - as.Date("2020-01-01")
case_date$cum_case <- cumsum(case_date$n)
```

```
plot(case_date$date,case_date$n,xlab="Case report date to CDC (0 as first date 2020-01-01)",ylab="Covid
```



There is a moderate positive linear relationship between CDC report date and Covid case number.

```
plot(case_date$date,case_date$cum_case,xlab="Case report date to CDC (0 as first date 2020-01-01)",ylab=
```



There is a strong positive nonlinear relationship between CDC report date and culmulative Covid case number.

**Appendix Data:**

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

```
data_dics <- data.frame("Column names"=NA, "Description"=NA, "Type"=NA)
data_dics[1,] = c("cdc_report_dt", "Initial case report date to CDC", "Quantative: Date & Time")
data_dics[2,] = c("pos_spec_dt", "Date of first positive specimen collection", "Quantative: Date & Time
data_dics[3,] = c("onset_dt", "Symptom onset date, if symptomatic", "Quantative: Date & Time")
data_dics[4,] = c("current_status", "Case Status: Laboratory-confirmed case; Probable case", "Categorica
data_dics[5,] = c("sex", "Sex: Male; Female", "Categorical: Text")
data_dics[6,] = c("age_group", "Age Group: 0 - 9 Years; 10 - 19 Years; 20 - 39 Years; 40 - 49 Years; 50
80 + Years", "Categorical: Text")
data_dics[7,] = c("Race and ethnicity (combined)", "Race and ethnicity (combined): Hispanic/Latino; Ame
data_dics[8,] = c("hosp_yn", "Hospitalization status: Yes/No", "Categorical: Text")
data_dics[9,] = c("icu_yn", "ICU admission status: Yes/No", "Categorical: Text")
data_dics[10,] = c("death_yn", "Death status: Yes/No ", "Categorical: Text")
data_dics[11,] = c("medcond_yn", "Presence of underlying comorbidity or disease: Yes/No", "Categorical:
knitr::kable(data_dics)
```

| Column.names | Description | Type |
|---|---|---|
| cdc_report_dt | Initial case report date to CDC | Quantative: Date & Time |
| pos_spec_dt | Date of first positive specimen collection | Quantative: Date & Time |

| Column.name | Description | Type |
|---|---|---|
| onset_dt | Symptom onset date, if symptomatic | Quantative: Date & Time |
| current_status | Case Status: Laboratory-confirmed case; Probable case | Categorical: Text |
| sex | Sex: Male; Female | Categorical: Text |
| age_group | Age Group: 0 - 9 Years; 10 - 19 Years; 20 - 39 Years; 40 - 49 Years; 50 - 59 Years; 60 - 69 Years; 70 - 79 Years; | |
| 80 + Years | Categorical: Text | |
| Race and ethnicity (combined) | Race and ethnicity (combined): Hispanic/Latino; American Indian / Alaska Native, Non-Hispanic; Asian, Non-Hispanic; Black, Non-Hispanic; Native Hawaiian / Other Pacific Islander, Non-Hispanic; White, Non-Hispanic; Multiple/Other, Non-Hispanic | Categorical: Text |
| hosp_yn | Hospitalization status: Yes/No | Categorical: Text |
| icu_yn | ICU admission status: Yes/No | Categorical: Text |
| death_yn | Death status: Yes/No | Categorical: Text |
| medcond_yn | Presence of underlying comorbidity or disease: Yes/No | Categorical: Text |

‘cdc_report_dt’: Initial case report date to CDC (Date & Time)

‘pos_spec_dt’: Date of first positive specimen collection (Date & Time)

‘onset_dt’: Symptom onset date, if symptomatic (Date & Time)

‘current_status’: Case Status: Laboratory-confirmed case; Probable case (Text)

‘sex’: Sex: Male; Female; Unknown; Other (Text)

‘age_group’: Age Group: 0 - 9 Years; 10 - 19 Years; 20 - 39 Years; 40 - 49 Years; 50 - 59 Years; 60 - 69 Years; 70 - 79 Years; 80 + Years (Text)

‘Race and ethnicity (combined)’: Race and ethnicity (combined): Hispanic/Latino; American Indian / Alaska Native, Non-Hispanic; Asian, Non-Hispanic; Black, Non-Hispanic; Native Hawaiian / Other Pacific Islander, Non-Hispanic; White, Non-Hispanic; Multiple/Other, Non-Hispanic (Text)

‘hosp_yn’: Hospitalization status: Yes/No (Text)

‘icu_yn’: ICU admission status: Yes/No (Text)

‘death_yn’: Death status: Yes/No (Text)

‘medcond_yn’: Presence of underlying comorbidity or disease: Yes/No (Text)

Citation:

[1] Q&A on coronaviruses (COVID-19) https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses

[2] “COVIDView: A Weekly Surveillance Summary of U.S. COVID-19 Activity,” Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html. [Accessed: 28-Sep-2020].

[3] Lewis NM, Friedrichs M, Wagstaff S, et al. Disparities in COVID-19 Incidence, Hospitalizations, and Testing, by Area-Level Deprivation — Utah, March 3–July 9, 2020. MMWR Morb Mortal Wkly Rep 2020;69:1369–1373. DOI: http://dx.doi.org/10.15585/mmwr.mm6938a4external icon

[4] Boehmer TK, DeVies J, Caruso E, et al. Changing Age Distribution of the COVID-19 Pandemic — United States, May–August 2020. MMWR Morb Mortal Wkly Rep. ePub: 23 September 2020. DOI: http://dx.doi.org/10.15585/mmwr.mm6939e1external icon.