Comparing Texas COVID-19 Data To Other US States With High Mortality and Population Density

Leeza Sergeeva

December 6, 2020

Background and Overview

This report consists of two parts.

First part uses user input for any date to gain insight of situation in the United States during the COVID-19 pandemic. By providing an argument in the form of a date (YYYYMMDD), we can look at total death numbers, number of positive cases, and number of total hospitalizations up to the chosen date and compare those to Texas COVID-19 statistics. This data is also compared with the 2020 United States population density and the general time line of mortality increase and the amount on positive tests increase for the top 5 states with the highest number of deaths up to chosen date, including Texas. The states with the highest population have the highest number of deaths.

Second part consists of variant analysis of SARS-CoV-2 sequencing data from Texas obtained from BioProject on NCBI website. This was an adaptation of the previous SARS-CoV-2 variant analysis (Koyama *et al.*, 2020). As predicted S and N genes have the highest number of SNPs variants in analyzed Texas samples possibly due to their larger size comparing to the other genes.

Methods

COVID-19 data

The data is provided in real time using API from the COVID Tracking Project website (Covid-19 tracking data api, 2020). The following columns were selected for the visualization: 'date' column represents date as YYYYMMDD on which data was collected by The COVID Tracking Project. The earliest date that can be used is 20200122. 'state' column represents two-letter abbreviation for the state. 'death' column represents total fatalities with confirmed OR probable COVID-19 case diagnosis. 'deathIncrease' column represents daily increase in death, calculated from the previous day's value. 'hospitalizedCumulative' column represents total number of individuals who have ever been hospitalized with COVID-19. 'hospitalizedIncrease' column represents daily increase in hospitalizedCumulative, calculated from the previous day's value. 'positive' column represents total number of confirmed plus probable cases of COVID-19 'positiveIncrease' column represents the daily increase positive cases (confirmed plus probable) calculated based on the previous day's value.

US Population Density Data

The US population density data was downloaded from this website as a .csv table. Using Excel the 'State' column was modified to match state's two letter codes.

SARS-CoV-2 Sequences

On November 16, 2020 I downloaded "TX SARS-CoV-2 Sequencing" SraRunTable from BioProject on NCBI website related to SARS-CoV-2. Here are the links to "TX SARS-CoV-2 Sequencing" BioProject: "https://www.ncbi.nlm.nih.gov/bioproject/PRJNA639066".

Reference SARS-CoV-2 isolate Wuhan-Hu-1, complete genome sequence was downloaded on October 15, 2020 from NCBI website "https://www.ncbi.nlm.nih.gov/nuccore/NC_045512".

Sequence data processing and filtering, from .fastq to .vcf files

Through series of executed shell scripts the .fastq files referenced in SraRunTables_tx.txt file were downloaded to external server provide by the University of San Francisco, as well as the SARS-CoV-2 reference genome file, the annotation gff for SARS-CoV-2. Then FastQC program processed .fastq files downloaded to validate the quality of the high throughput sequencing data sets (Andrews *et al.*, 2012). Next step used trimmomatic function to clean up sequencing data by throwing out bad sequences. With the help of Burrows-Wheeler Alignment Tool reference genome was indexed, and the reads in each of the samples provided were aligned to the reference genome and saved as .sam files (Li and Durbin, 2009). Using Samtools, .sam files were converted to .bam files and sorted by leftmost coordinates. Then each sorted .bam file was processed by Samtools flagstat to count number of alignments in each FLAG type, like QC pass, QC fail etc (Li *et al.*, 2009). Using bcftools mpileup on sorted .bam files generate .bcf file that contains genotype likelihoods at each genomic position with coverage and then SNPs are called for each input file saving output of that as .vcf files and filtering out the short variants for the final VCF.

See the set of tutorials on the vcfR package website.

Results and Discussion

The US states with the highest populations in 2020 are: CA, TX, FL, NY, PA, IL @ref(tab:top15-population-density-table). The US states with the highest population density are: DC, NJ, RI, MI, CT @ref(tab:top15-total-population-table). States with the most cumulative deaths as of November 30, 2020 are: NY, TX, CA, FL, NJ @ref(fig:total-deaths-plot). This looks like the total population might be related to the total number of deaths related to COVID-19. As most of those states with high mortality are close to the ocean they contradict findings done in Japan, where lower mortality rates were observed for higher temperature and absolute humidity (Kodera *et al.*, 2020). Looking into humidity and average temperature data in the United States would provide more incite. States with most hospitalizations up to selected date are: NY, FL, NJ, GA, OH @ref(fig:total-hospital-plot). Note: there was no hospitalization data for Texas provided in the API. States with most positive test results up to selected date are: CA, TX, FL, IL, NY @ref(fig:total-positive-plot).

It seems like the population density is not directly related to the spread of COVID-19. It might be helpful to look at the data on county or city level to determine if the population density contributed to the spread of COVID-19. However, it's been shown that the spread of the disease has moderate association with population density based studies done on India's COVID-19 data (Bhadra *et al.*, 2020). Another study showed no statistically significant relationship between the spread of the disease and population density (Valev, 2020). As the number of positive tests increases, so does the number of dead people in the state @ref(fig:increase-death-positive-test-plots). This finding is similar to that presented in another paper advocating closest relationship of Deaths per million population and total Cases per million population (Valev, 2020). Based on total population similarities with California and New York, Texas's relationship to COVID-19 is very similar. It would be interesting to compare the humidity, and temperature data in the future.

The variant analysis of SARS-CoV-2 sequencing data from Texas shows highest number of SNPs variants in S (spike glycoprotein) and N (nucleocapsid phosphoprotein) genes @ref(fig:unique-SNPs-plot). This could be due to the relatively large size of those two genes comparing to the other ones @ref(tab:sars-cov-2-genes-table). This finding is consistent with the published variant analysis of SARS-CoV-2 genomes (Koyama *et al.*, 2020).

Figures

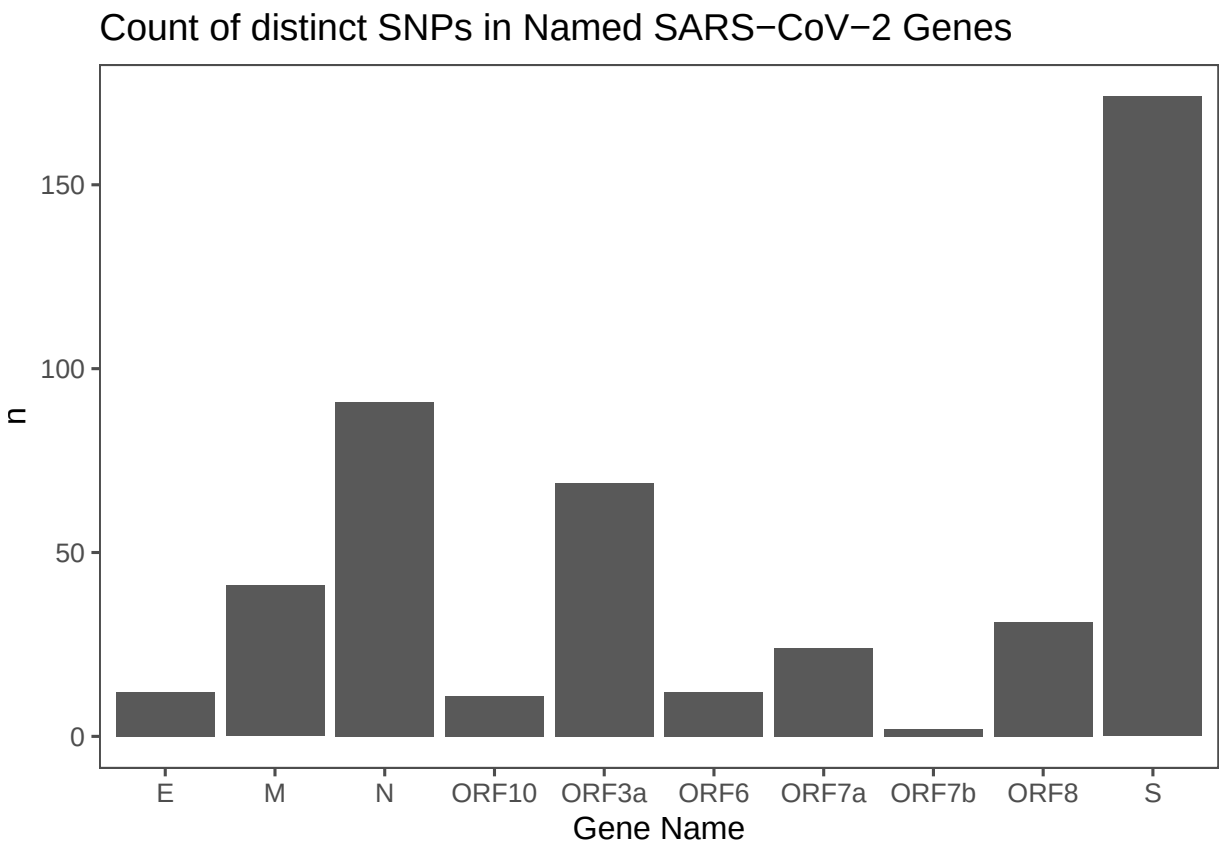

Figure 1: N and S genes have more unique SNPs in the set of samples analyzed.

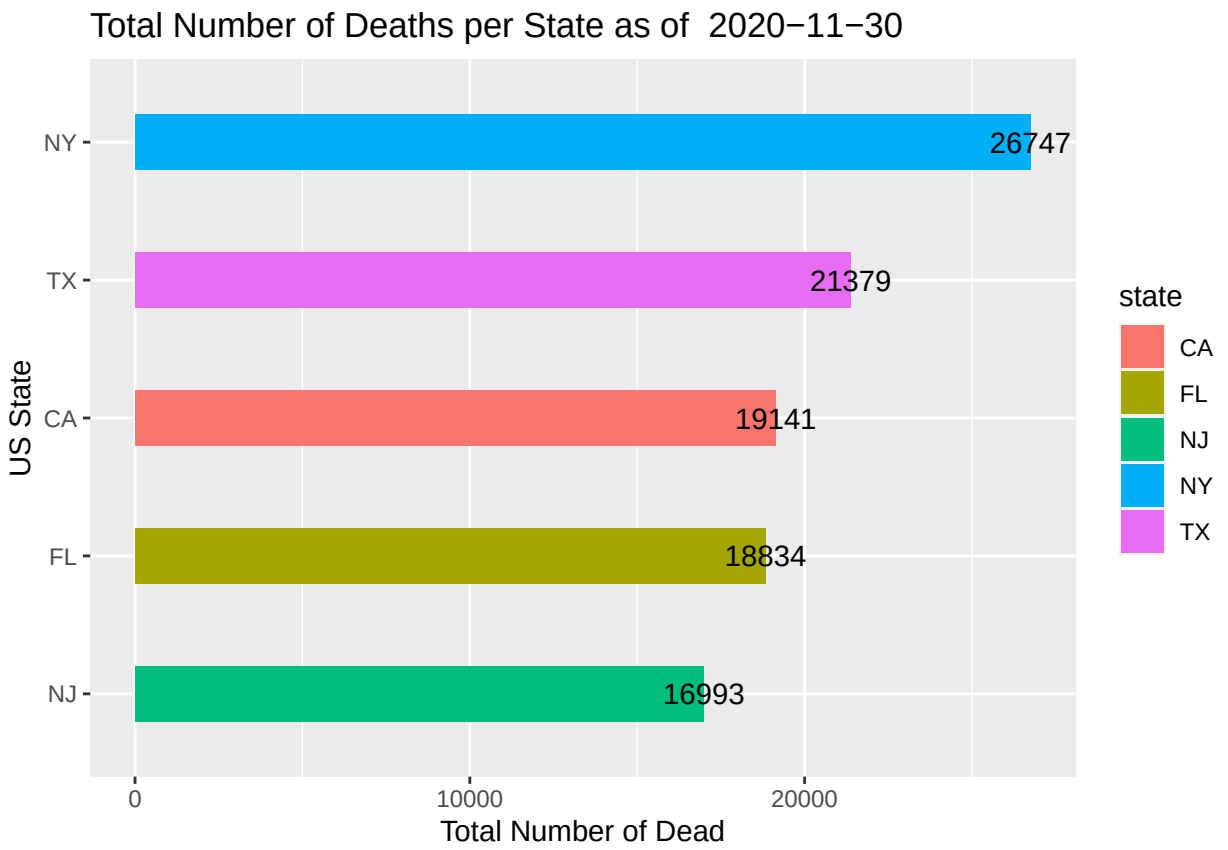

Figure 2: The bar plot of the top 5 US States (including Texas) with the highest total number of COVID-19 related deaths on given date.

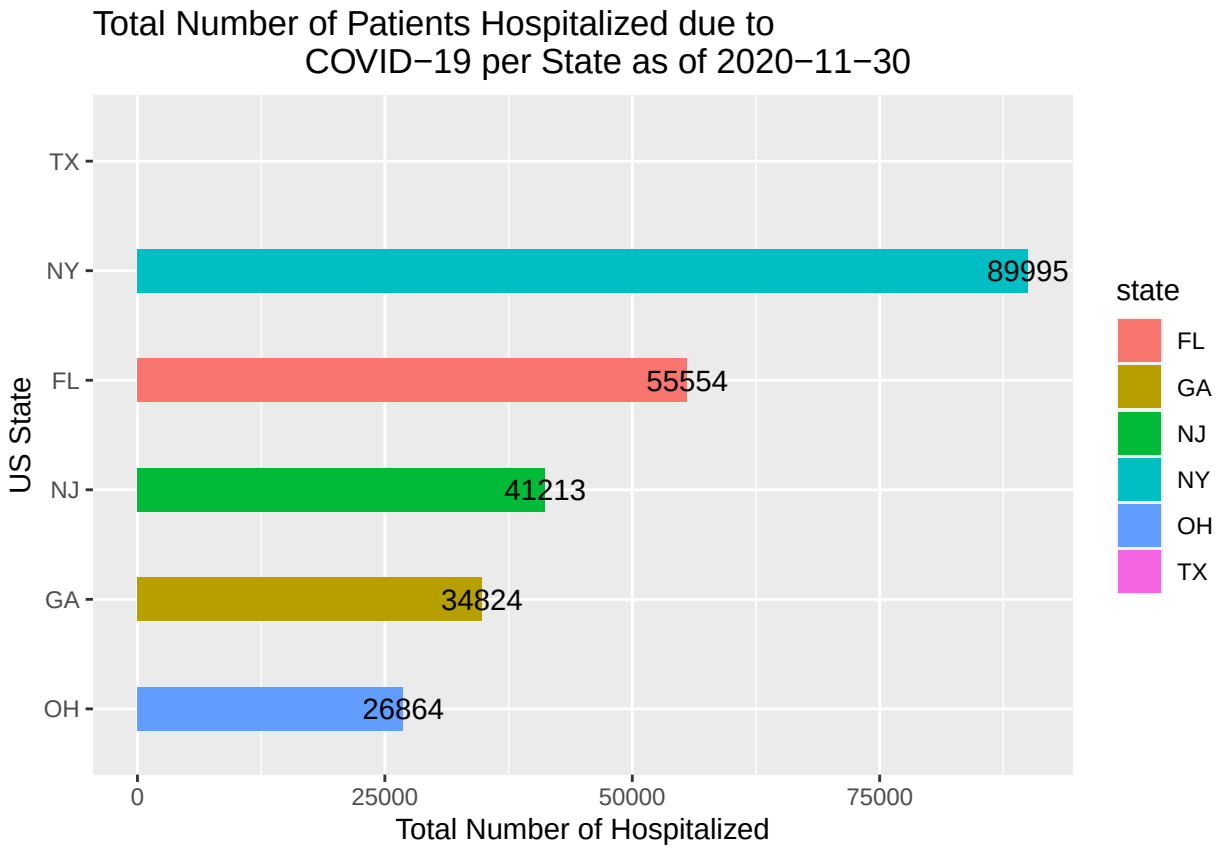

Figure 3: The bar plot of the top 5 US States (including Texas) with the highest cumulative number of COVID-19 related hospitalizations on given date. Note: there is no hospitalization data available for Texas.

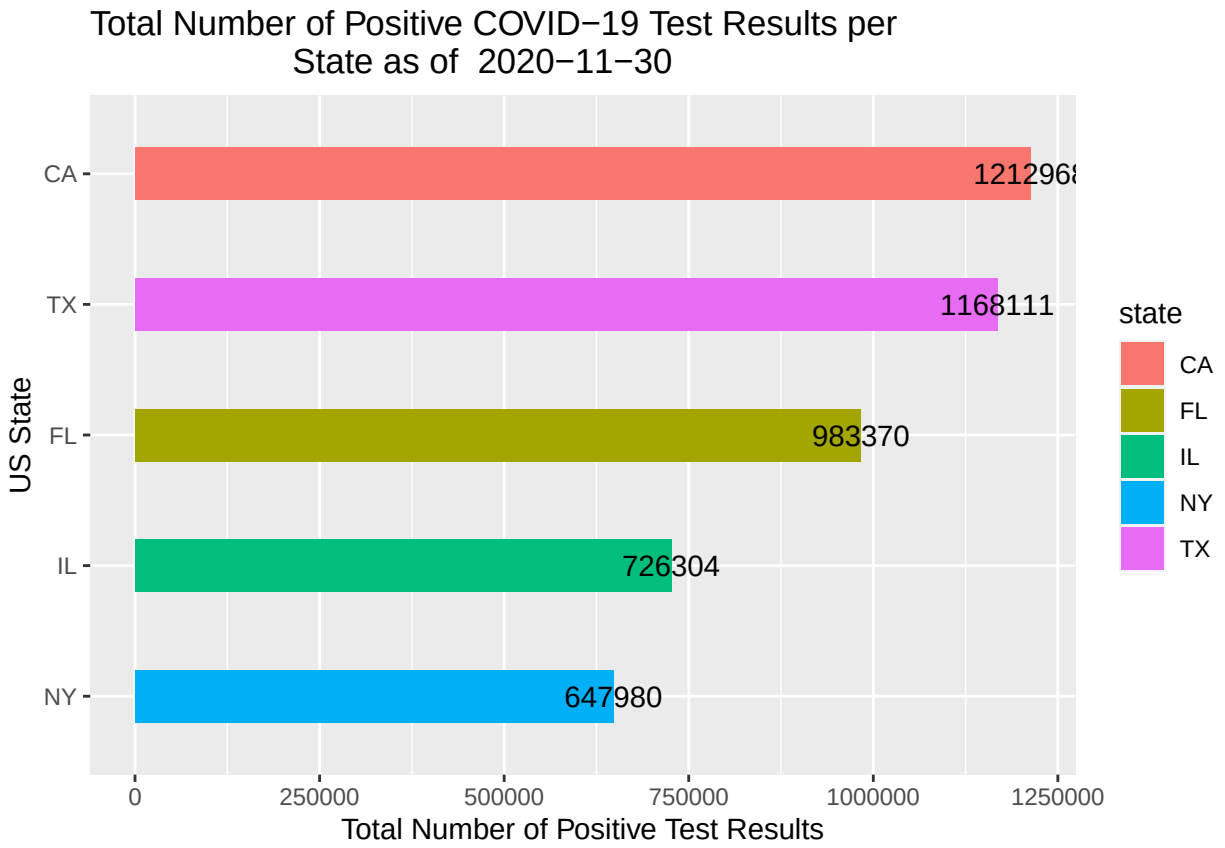

Figure 3: The bar plot of the top 5 US States (including Texas) with the highest cumulative number of COVID-19 positive test results.

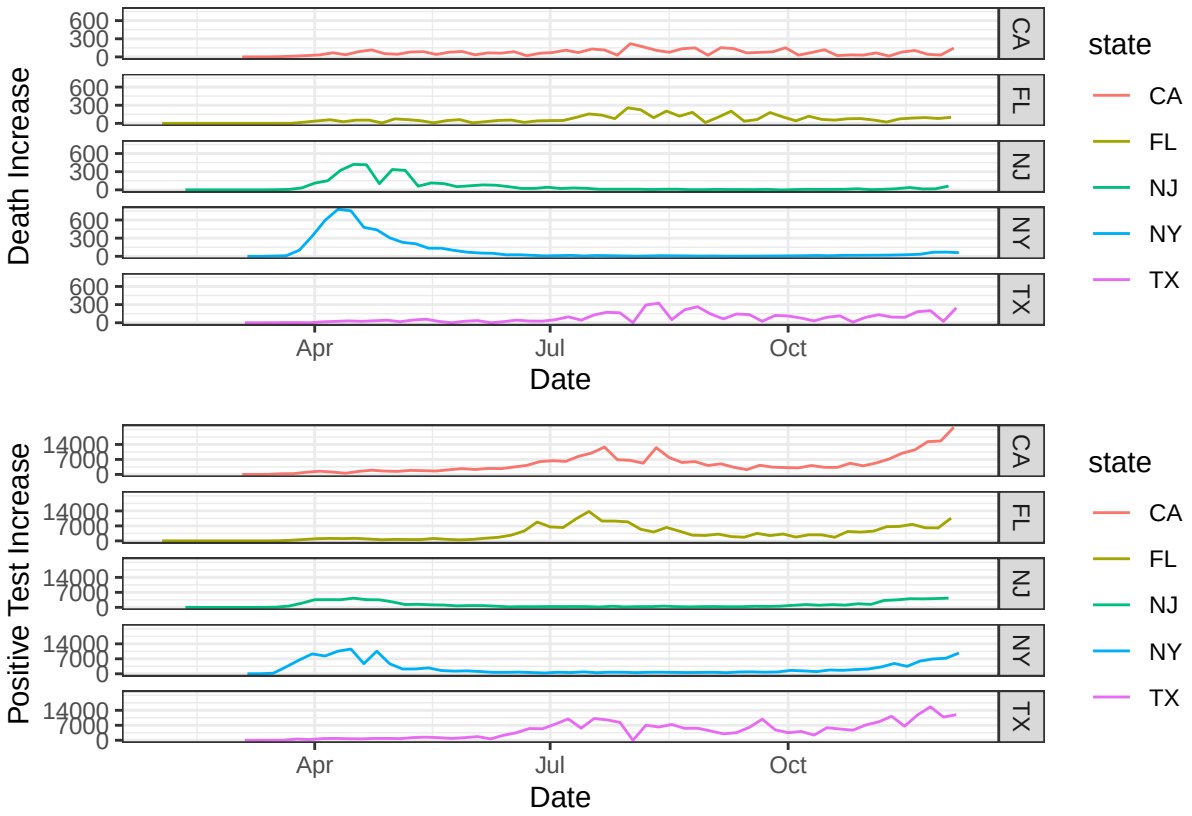

Figure 4: Top graph is a time series plots of mortality increase and bottom graph is a time series of COVID-19 positive tests increase for the top 5 US States (including Texas) with the highest cumulative number of COVID-19 related deaths on given date.

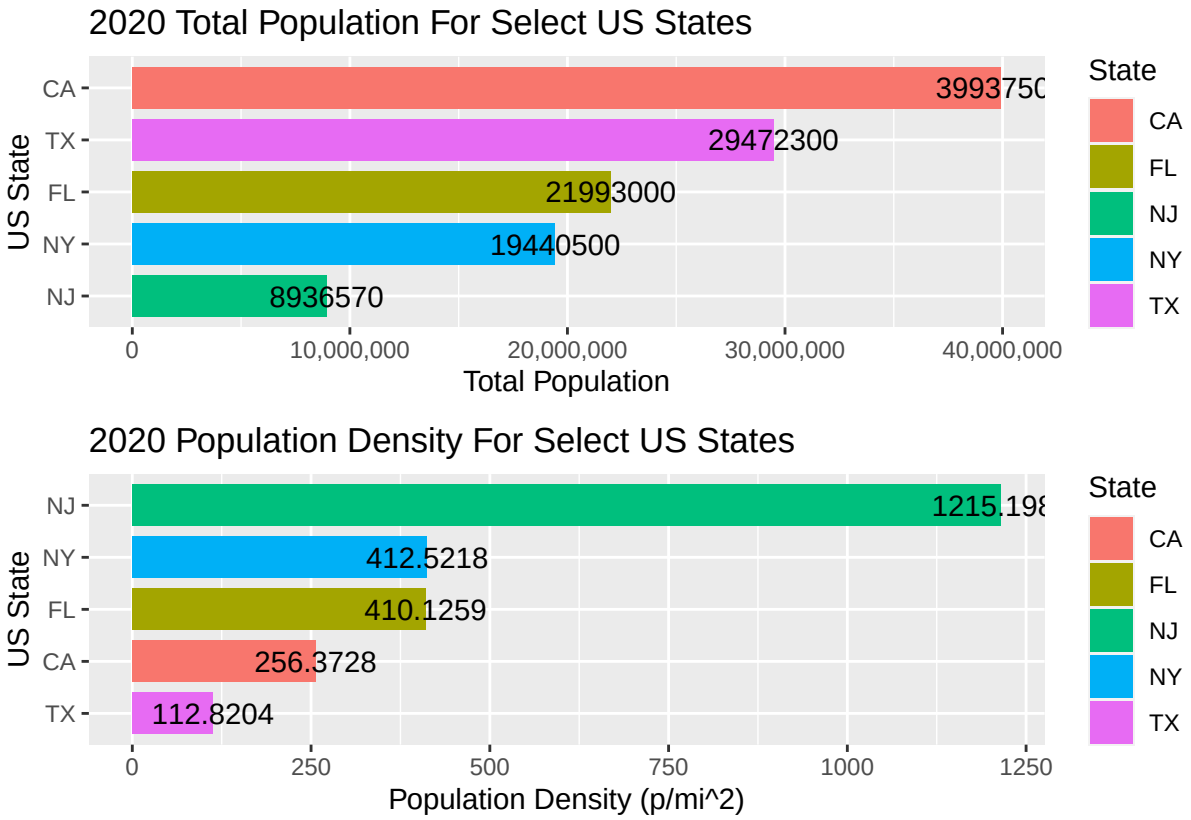

Figure 5: Top graph shows total US population in 2020 for select 5 states that have high number of deaths up to select date. Bottom graph shows population density in 2020 of the related 5 states with the high number of deaths up to selected date.

Tables

Table 1: SARS-CoV-2 Gene Names, Locations, and Lengths

Gene Name	Start	End	Length
S	21563	25384	3821
ORF3a	25393	26220	827
E	26245	26472	227
M	26523	27191	668
ORF6	27202	27387	185
ORF7a	27394	27759	365
ORF7b	27756	27887	131
ORF8	27894	28259	365
N	28274	29533	1259
ORF10	29558	29674	116

Table 1: Gene names, locations, and lengths in the SARS-CoV-2 genome. Higher SNP counts in the S and N genes may be related to the larger size of these genes.

Table 2: Top 15 US States With Highest Population Density in 2020

State	Pop Density (p/mi^2)
DC	11814.5410
NJ	1215.1985
RI	1021.4313
MA	894.4359
CT	735.8695
MD	626.6735
DE	504.3073
NY	412.5218
FL	410.1259
OH	287.5040
PA	286.5454
CA	256.3728
IL	228.0246
HI	219.9424
VA	218.4404

Table 2: Top 15 states with highest population densities in 2020.

Table 3: Top 15 US States With Highest Population in 2020

State	Population
CA	39937500
TX	29472300
FL	21993000
NY	19440500
PA	12820900
IL	12659700
OH	11747700
GA	10736100
NC	10611900
MI	10045000
NJ	8936570
VA	8626210
WA	7797100
AZ	7378490
MA	6976600

Table 3: Top 15 states with highest total population in 2020.

Sources Cited

Andrews,S. *et al.* (2012) FastQC.

Bhadra,A. *et al.* (2020) Impact of population density on Covid-19 infected and mortality rate in India. *Model Earth Syst Environ*, 1–7.

Covid-19 tracking data api (2020) *The COVID Tracking Project*.

Kodera,S. *et al.* (2020) Correlation between COVID-19 Morbidity and Mortality Rates in Japan and Local Population Density, Temperature, and Absolute Humidity. *Int J Environ Res Public Health*, **17**.

Koyama,T. *et al.* (2020) Variant analysis of sars-cov-2 genomes. *Bulletin of the World Health Organization*, **98**, 495.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Valev,D. (2020) Relationships of total covid-19 cases and deaths with ten demographic, economic and social indicators. *medRxiv*.