IE 6600 Project Presentation

Analyzing and Predicting the causes and new cases for COVID-19 in the United States

Bananas: I-Hsuan Huang, Yiming Wang, Yen-Fong Li, Wenzheng Liao

Overview	Data Preprocessing	EDA	ML models
Introduction of the Project	Retrieve, Access, Clean and Transform Data.	Correlation Analytics Significance Testing Descriptive Analytics Visual Analytics	Linear Regression, Random Forest, ARIMA model

Overview

In this project, we utilize the CDC death and COVID-19 datasets to study the types of diseases affecting people's health and their relations to COVID-19, as well as to visualize interesting patterns and generate predictive time series models. Beginning by conducting an exploratory analysis of the datasets, we study the general data tidiness and plot distributions and correlation among multiple variables.

The goal of our study is to help the disease experts, epidemiologist researchers and the general public to get a deeper understanding of how COVID-19 affects the public health and who are most vulnerable to it.

Step 1: Retrieve Data

Way 1: Manually download

- Counts of Deaths for Select Causes of Death by Sex, Age, and Race and Hispanic Origin
- Conditions Contributing to COVID-19 Deaths, by State and Age
- United States COVID-19 Cases and Deaths by State over Time

Way 2: Via API

- API's doc

Step 2: Access Table 1

- Categorical Columns:
  - Sex, Race, Age Group
- Date Columns:
  - year of death, month of death
- Numerical Columns:
  - Count of death of a specific disease(including COVID-19)
- Example:
  - The value pointed by arrow represents for the **total number of people killed by Septicemia** who satisfies the following condition:
    - sex is **female**
    - race is **Other**
    - age group is **5-14 years**
    - die in **2020 March**

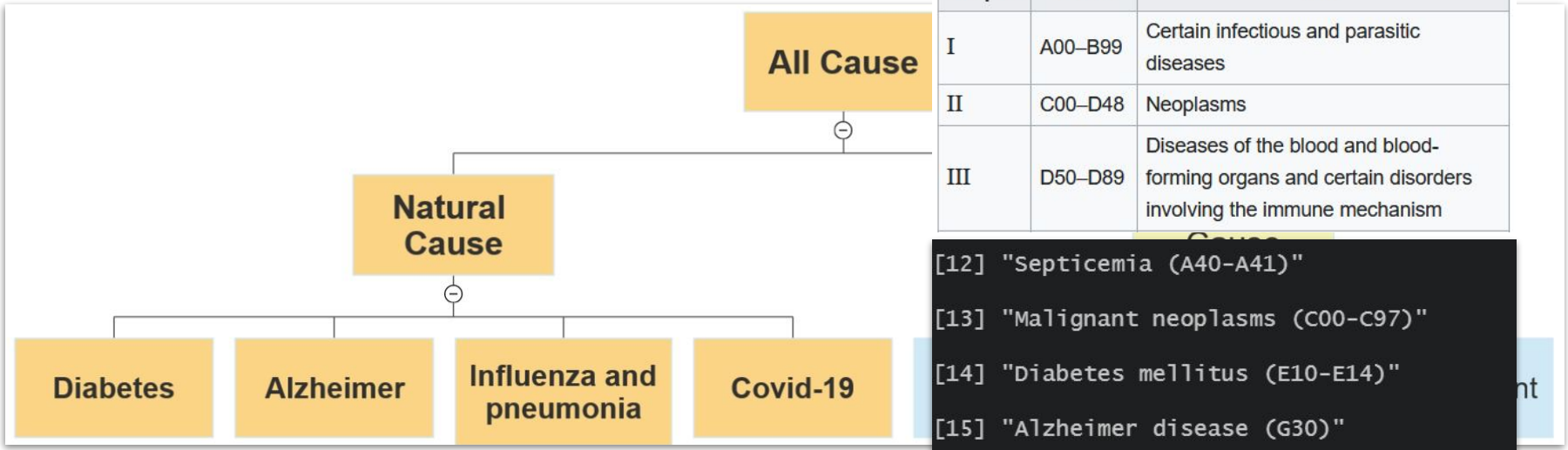
	Date Columns		Categorical Columns			Numerical Columns			
	year_death	month_death	sex	race	age_group	cause_natural	cause_other	Septicemia (A40-A41)	Malignant neoplasms (C00-C97)
1	2019	7	male	Other	0-4 years	52	9	0	1
2	2019	9	female	Other	25-34 years	8	18	0	1
3	2020	3	female	Other	0-4 years	35	5	0	0
4	2020	3	female	Other	5-14 years	4	2	1	0
5	2020	3	female	Other	15-24 years	2	12	0	0

Step 2: Important Note on Table 1

- Some Numerical Columns:
  - All Cause
  - Natural Cause
- Some medical terms:
  - ICD-10 code
  - COVID-19 Underlying Cause
  - COVID-19 Multiple Cause
- What are these ?
- What's there relation ?

```
> count_death %>% colnames()
 [1] "AnalysisDate"
 [2] "Date Of Death Year"
 [3] "Date Of Death Month"
 [4] "Start Date"
 [5] "End Date"
 [6] "Jurisdiction of Occurrence"
 [7] "Sex"
 [8] "Race/Ethnicity"
 [9] "AgeGroup"
[10] "AllCause"
[11] "NaturalCause"
[12] "Septicemia (A40-A41)"
[13] "Malignant neoplasms (C00-C97)"
[14] "Diabetes mellitus (E10-E14)"
[15] "Alzheimer disease (G30)"
[16] "Influenza and pneumonia (J09-J18)"
[17] "Chronic lower respiratory diseases (J40-J47)"
[18] "Other diseases of respiratory system (J00-J06,J30-J39,J67,J70-J98)"
[19] "Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27)"
[20] "Symptoms, signs and abnormal clinical and laboratory findings, not else
[21] "Diseases of heart (I00-I09,I11,I13,I20-I51)"
[22] "Cerebrovascular diseases (I60-I69)"
[23] "COVID-19 (U071, Multiple Cause of Death)"
[24] "COVID-19 (U071, Underlying Cause of Death)"
```

Illustration of Medical Terms: All Cause, Natural Cause, ICD-10 Code



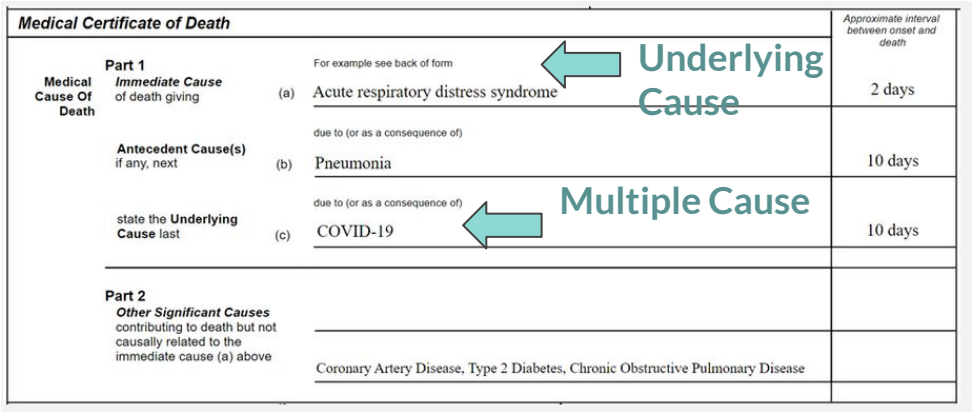
ICD-10 chapters

Chapter	Block	Title
I	A00–B99	Certain infectious and parasitic diseases
II	C00–D48	Neoplasms
III	D50–D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

```
[12] "Septicemia (A40-A41)"
[13] "Malignant neoplasms (C00-C97)"
[14] "Diabetes mellitus (E10-E14)"
[15] "Alzheimer disease (G30)"
[16] "Influenza and pneumonia (J09-J18)"
```

Step 2: Important Note on Table 1 con't

- Two similar columns:
  - **COVID-19 Underlying Cause**
  - **COVID-19 Multiple Cause**
- Our Decision
  - Keep only **COVID-19 Underlying Cause**



Sample of Death Certificate

source: https://cpsa.ca/news/physician-notes-basic-principles-on-medical-death-certification/

Step 2: Access Table 2

- Categorical Columns:
  - State, Condition, Condition Group, Age Group
- Date Columns:
  - Year, Month
- Numerical Columns:
  - COVID-19 Deaths
- Example:
  - The value pointed by arrow represents for the **total number of people killed by COVID-19** who satisfies the following condition:
    - state is **Iowa**
    - condition group is **Respiratory disease**
    - condition is **Influenza and Pneumonia**
    - die in **April** of **2020**

	year	month	state	condition_group	condition	age_group	covi_death
750	2020	1	Iowa	Respiratory diseases	Influenza and pneumonia	0-24	0
751	2020	2	Iowa	Respiratory diseases	Influenza and pneumonia	0-24	0
752	2020	3	Iowa	Respiratory diseases	Influenza and pneumonia	0-24	0
753	2020	4	Iowa	Respiratory diseases	Influenza and pneumonia	0-24	0
754	2020	6	Iowa	Respiratory diseases	Influenza and pneumonia	0-24	0

Date Columns

Categorical Columns

Numerical Columns

Step 2: Important Note on Table 2

- Counter-intuitive: Column State
  - Strange values:
    - United States
    - New York City and New York State
- Column **Group**
  - 3 distinct value:By Total, By Year, By Month
- Column ‘Condition Group’ and ‘Condition’
  - Each categorical value in column ‘Condition Group’ corresponds to several categorical values in column ‘Condition’
  - In conclusion, values in column ‘Condition Group’ is the category of values in column ‘Condition’

Investigate Column ‘Condition Group’ and ‘Condition’

```
> condition_covi %>%
+   filter(., `Condition Group`=='Respiratory diseases') %>%
+   distinct(., Condition)
# A tibble: 6 × 1
  Condition
  <chr>
1 Influenza and pneumonia
2 Chronic lower respiratory diseases
3 Adult respiratory distress syndrome
4 Respiratory failure
5 Respiratory arrest
6 Other diseases of the respiratory system
```

Step 2: Access Data 3

- Categorical Columns:
  - state
- Date Columns:
  - Date: 1001 dates
- Numerical Columns:
  - total cases, new cases, total death, new death
  - other numerical columns are dropped, like probable cases
- Example:
  - The value annotated by arrow represents for the **total number of people who tested positive for COVID-19** who satisfy the following conditions:
    - it is tested positive on **26 April 2021**
    - the state is **Maryland**

	date	state	tot_cases	new_case	tot_death	new_death	conf_cases	prob_cases	pnew_case	conf_death	prob_death	pnew_death
22480	2021-04-23	Maryland	441155	1163	9230	11	NA	NA	0	9018	212	0
22481	2021-04-24	Maryland	442351	1196	9248	18	NA	NA	0	9036	212	0
22482	2021-04-25	Maryland	443257	906	9263	15	NA	NA	0	9049	214	2
22483	2021-04-26	Maryland	443814	557	9274	11	NA	NA	0	9060	214	0
22484	2021-04-27	Maryland	444491	677	9296	22	NA	NA	0	9082	214	0

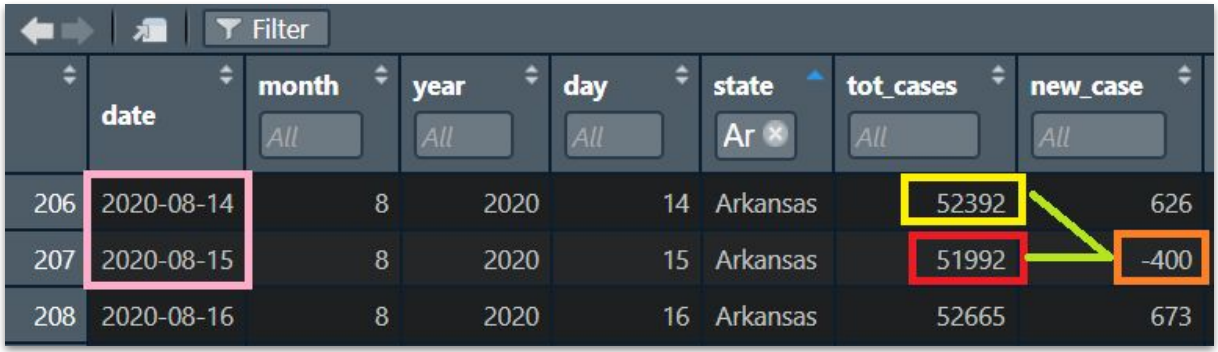
Numerical Columns

Step 2: Important Note on Table 3

- Counter-intuitive: Column State
  - 60 Unique Values
    - 51 states and DC
    - 1 city: The New York City
    - 7 US's overseas territories: U.S. territories of American Samoa, Guam, the Commonwealth of the Northern Mariana Islands, Puerto Rico, and the U.S Virgin Islands
- issue
  - Is ‘RMI’ and ‘RI’ a data issue ?
    - No. ‘RMI’ stands for Rhode Island, and ‘RI’ stand for Republic of Marshall Island.

Step 2: Important Note on Table 3 con't

- Negative Values(in orange)
- Verification:
  - 52392(yellow)+(-400)(orange)=51992(red)
- Conclusion:
  - Some states fix their data reported.
  - No cleaning required.



Filter

	date	month	year	day	state	tot_cases	new_case
		All	All	All	Ar	All	All
206	2020-08-14	8	2020	14	Arkansas	52392	626
207	2020-08-15	8	2020	15	Arkansas	51992	-400
208	2020-08-16	8	2020	16	Arkansas	52665	673

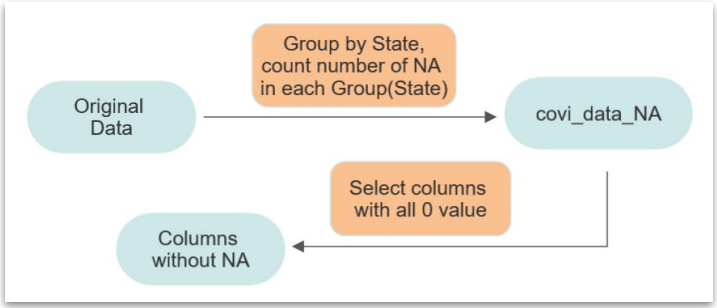
Step 2: Important Note on Table 3 con't

- NA Values in Numerical Columns
- 3 Conditions:
  - no NA values for all states(shown in green)
    - total cases, new cases, total death, new death
  - Some states have 1001 NA values in a column(shown in red)
  - Some states have NA values before a date and no NA values after(shown in yellow)
- Conclusion:
  - Not Data Tidiness issue
  - The reason is different states report their data in different ways.

covi_data_NA

	state	tot_cases_NA	conf_cases_NA	prob_cases_NA	new_case_NA
1	AK	0	1001	1001	0
2	AL	0	0	0	0
3	AR	0	1001	1001	0
4	AS	0	1001	1001	0
5	AZ	0	88	88	0
6	CA	0	0	0	0
7	CO	0	0	0	0
8	CT	0	87	87	0
9	DC	0	1001	1001	0
10	DE	0	0	0	0

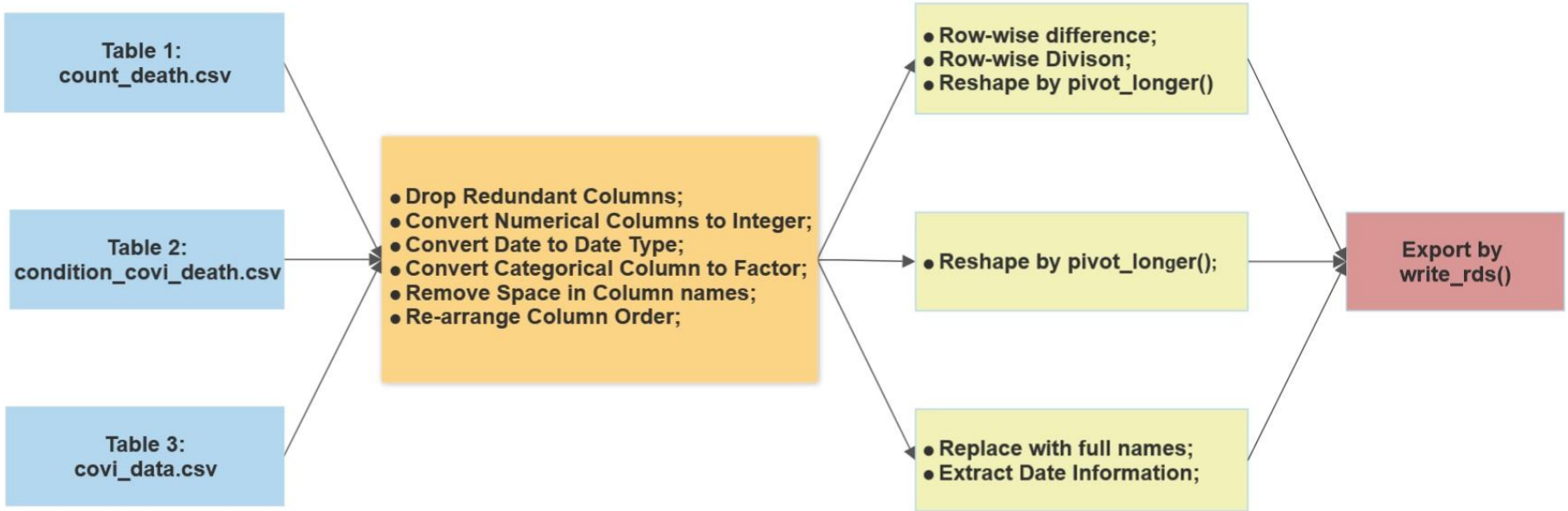
Process of Exploring NA values



Columns without NA

```
> # Which column has no NA-value among all states ?
> covi_data_NA %>%
+   select(.,
+          where(~ is.numeric(.x) && sum(.x) == 0))
# A tibble: 60 × 4
  tot_cases_NA new_case_NA tot_death_NA new_death_NA
         <int>       <int>        <int>        <int>
1            0           0            0            0
```

Step 3 and 4: Cleaning & Transformation & Export



Some Functions used

- A custom function
  - Replace abbr with full name
  - Input: string of abbr.
  - Output: string of full name
- apply function inside mutate()
  - for column-wise operation, it should be used inside sapply()

```
Now apply the custom function to covi_data_c1.csv
```{r}
# note that the name attributes should be removed
covi_data_c2 = covi_data_c1 %>%
  mutate(.,
         state = sapply(state, replace_abbr_fullname),
         state = unname(state))
```
```

```
# string vector of all abbr. of state
abbr_all_states =  c("AL","AK", "AZ", "KS", "UT", "CO", "CT",
                     "DE", "FL", "GA", "HI", "ID", "IL", "IN", "IA",
                     "AR", "KY", "LA", "ME", "MD", "MA", "MI", "MN",
                     "MS", "MO", "MT", "NE", "NV", "NH", "NJ", "NM",
                     "NY", "NC", "ND", "OH", "OK", "OR", "PA", "RI",
                     "SC", "SD", "TN", "TX", "CA", "VT", "VA", "WA",
                     "WV", "WI", "WY", "DC", "NYC",
                     "AS", "GU", "MP", "PR", "VI", "FSM", "PW", "RMI")

# string vector of all full name of state
full_all_states = c("Alabama","Alaska", "Arizona", "Kansas",
                    "Utah", "Colorado", "Connecticut",
                    "Delaware", "Florida", "Georgia",
                    "Hawaii", "Idaho", "Illinois",
                    "Indiana", "Iowa", "Arkansas",
                    "Kentucky", "Louisiana", "Maine",
                    "Maryland", "Massachusetts", "Michigan",
                    "Minnesota", "Mississippi", "Missouri",
                    "Montana", "Nebraska", "Nevada",
                    "New Hampshire", "New Jersey", "New Mexico",
                    "New York", "North Carolina", "North Dakota",
                    "Ohio", "Oklahoma", "Oregon",
                    "Pennsylvania", "Rhode Island", "South Carolina",
                    "South Dakota", "Tennessee", "Texas",
                    "California", "Vermont", "Virginia",
                    "Washington", "West Virginia", "Wisconsin",
                    "Wyoming", "District of Columbia", "New York City",
                    "American Samoa", "Guam", "Northern Mariana Islands",
                    "Puerto Rico", "U.S. Virgin Islands",
                    "Micronesia", "Palau", "Republic of Marshall Islands")

# define custom function to replace abbr. with full name
replace_abbr_fullname = function(abbr){
  str1 = tolower(abbr_all_states)
  str2 = full_all_states
  str2[match(tolower(abbr), str1)]
}
```

Some Useful tidyverse Functions

- across(): apply functions to multiple columns at one time

```
# convert date type
condition_covi_c2 = condition_covi_c1 %>%
  mutate(.,
         across(
           where(is.numeric), as.integer),
         across(
           c(State, `Condition Group`, Condition, ICD10_codes, `Age Group`), as.factor))
```

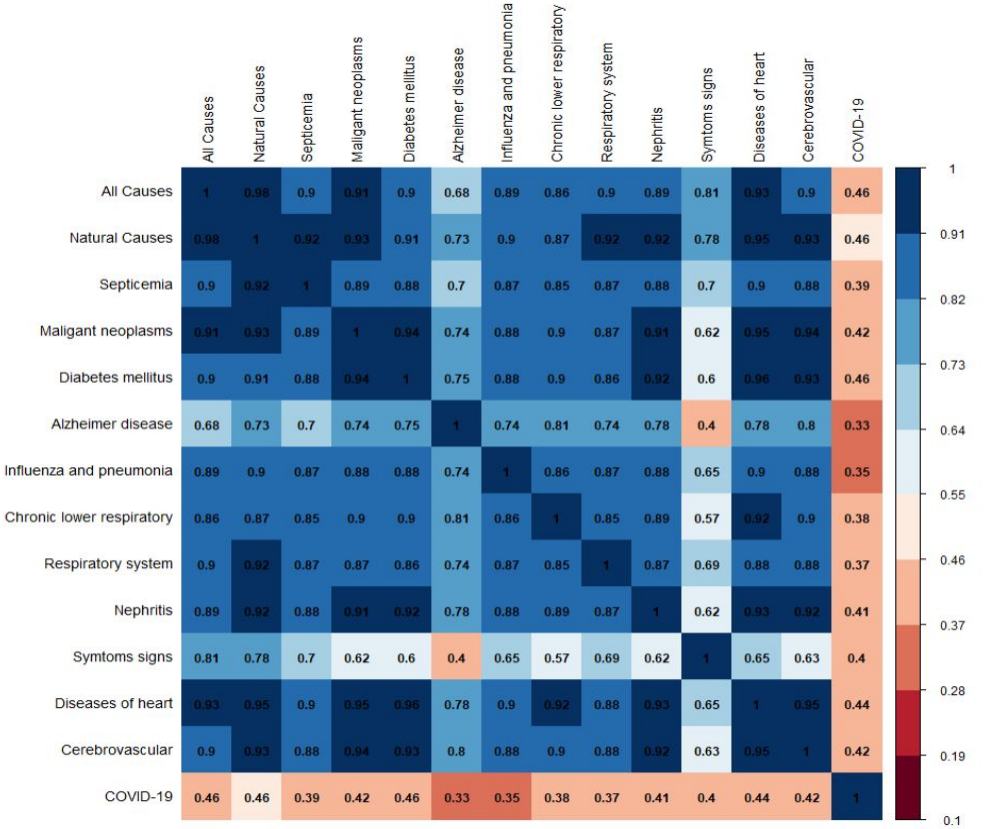
- fct_collapse(): merge factor levels

```
# merge old levels to new levels
count_death_c4 = count_death_c3 %>%
  mutate(.,
         Sex = fct_collapse(Sex,
                            male=c('M', 'Male'),
                            female=c('F', 'Female')))
```

- rename_with(): rename columns with functions as input

```
# change column names to lower-case, and replace space with underscore
condition_covi_c3 = condition_covi_c2 %>%
  rename_with(.,
              .fn = ~ tolower(gsub(" ", "_", .x, fixed = TRUE)),
              .cols = everything())
```

EDA: Correlation Analytics & Significance Test



Research Question:

Is there a statistically significant relationship between Covid 19's deaths and different underlying health conditions?

Research Method:

Spearman Correlation

Spearman Assumptions:

-data must be ordinal

- variables must be monotonically related to the other variables.

Key Takeaway:

- Most underlying health conditions have a **moderately** to **highly** correlated and statistically significant relationship ($0.33 \leq \rho \leq 0.96$, $P < 0.05$) with Covid-19 deaths .

- Disease of heart, Diabetes Mellitus, and respiratory systems have **positively correlated** relationships with the Covid 19 deaths.

Shiny Application Background

The **objective** of the Application is to provide an interactive analytics platform, allowing users to investigate patterns and trends of Covid-19 deaths and causes. It helps identify if there are any **geographic, demographic,** and/or **medical factors** that contribute to Covid-19 deaths.

Research Questions

- What are the primary underlying health conditions causing the most Covid-19 deaths?
- Which group(s) is at high risk from severe coronavirus diseases ?
- How does the Covid-19 death trend vary in different seasons and states?
- Which US states have the highest confirmed cases and death rates?

Solutions / Methodology

- Interactive Analytics Dashboard
- EDA Analytics: Correlation Analytics & Descriptive Analytics
- Time Series Forecasting & Modeling
- KPI Analytics

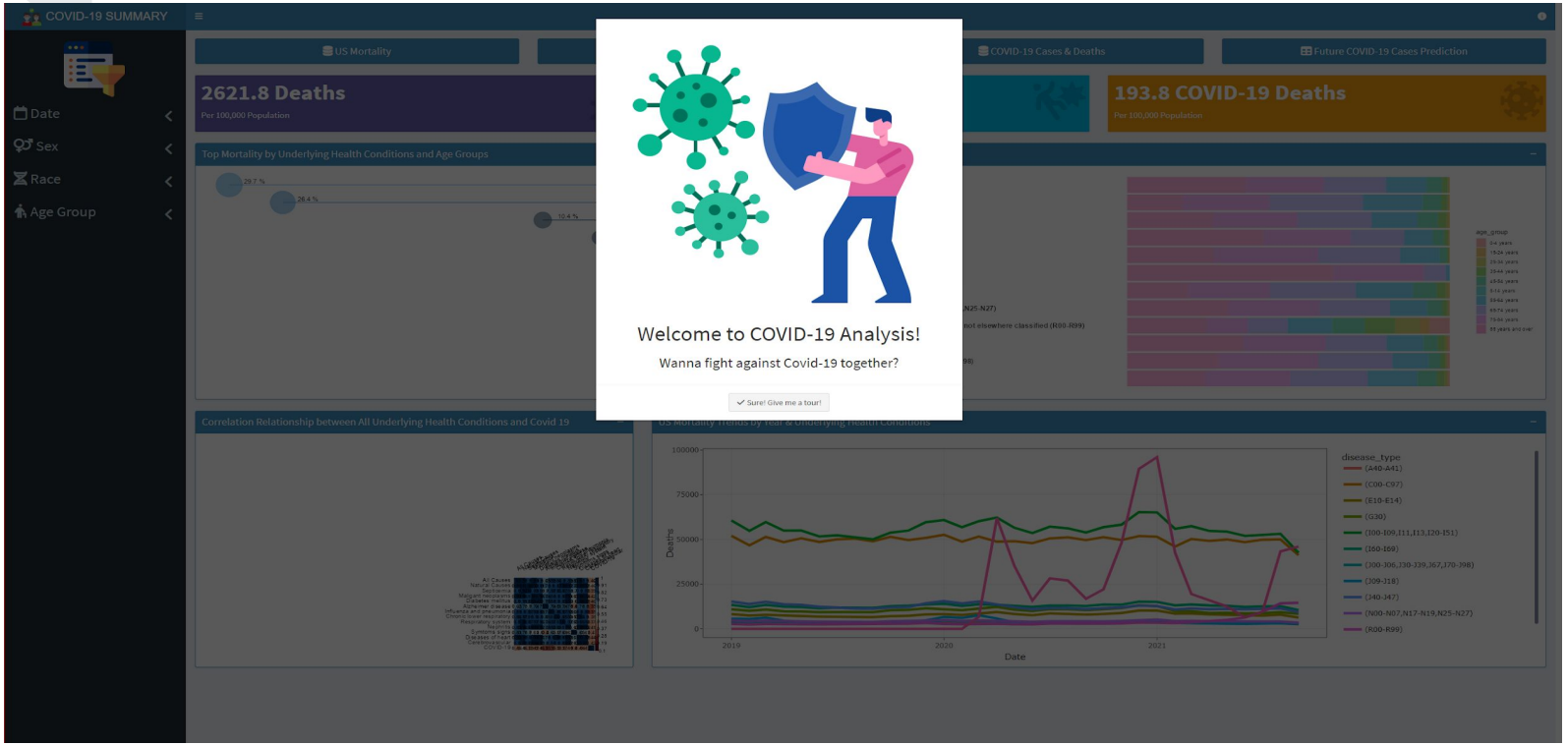
Key Findings

- **Disease of heart, Malignant neoplasms, and Covid-19** were the top underlying health conditions leading to US mortality between 2020-2022.
- **California, Texas, and Florida** were the three top states contributing to the highest Covid 19 death cases, as they were the top three states with the highest US population.
- **Disease of heart, diabetes Mellitus, and respiratory systems** have statistically significant relationships with the Covid 19 deaths.
- **December and January** (holiday season) were the two months with relatively higher confirmed and death cases, potentially due to the increased public crowds and gatherings.
- **Age groups of 65+** are the most vulnerable groups from Covid-19 deaths.
- **Predictions of future 100 days** of new COVID-10 cases generated using ARIMA model.

Research Challenges/Caveats

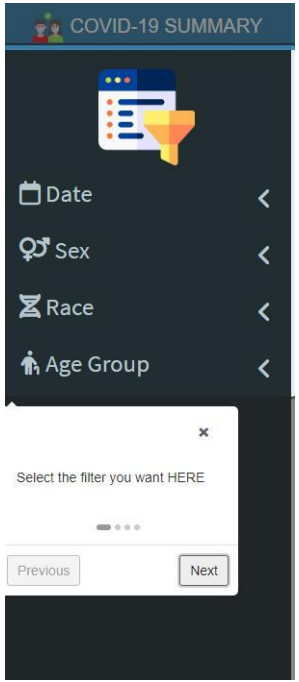
- **Vaccinations rates, mask wearing mandates by state government, quarantine rules** and other related factors did not take into consideration for this research.
- **External factors (Climate, air pollution, and the built environment)** were difficult to be considered in this study.

Two Interesting Shiny App Implementations

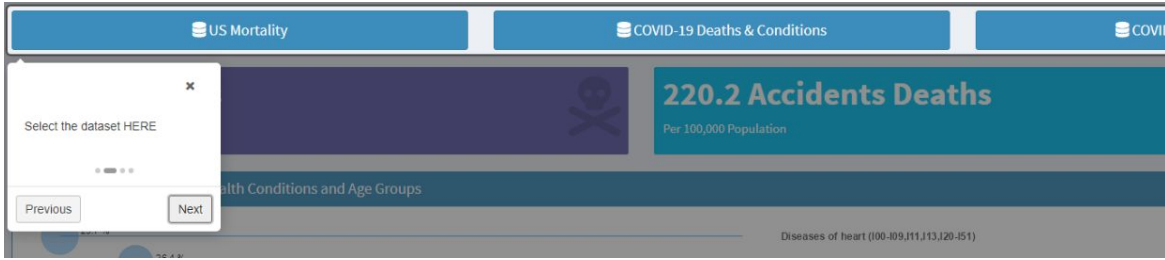


introjsUI
introBox()

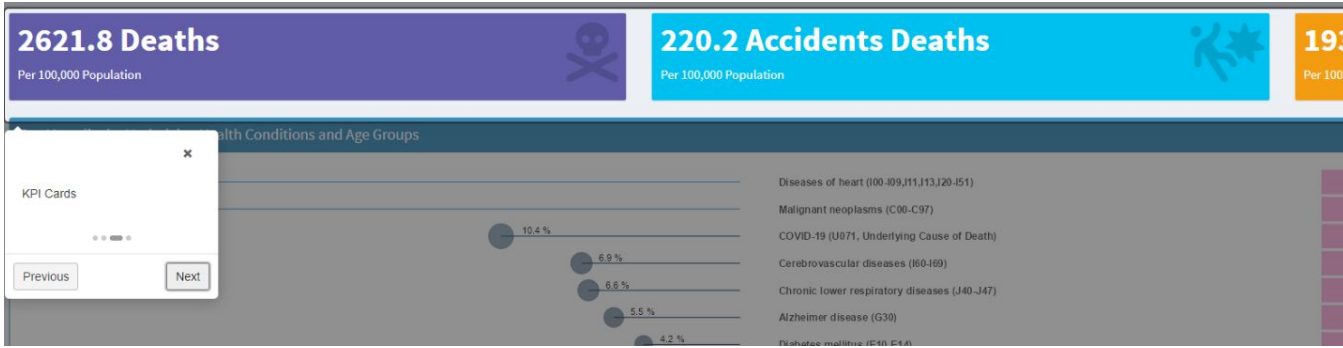
1



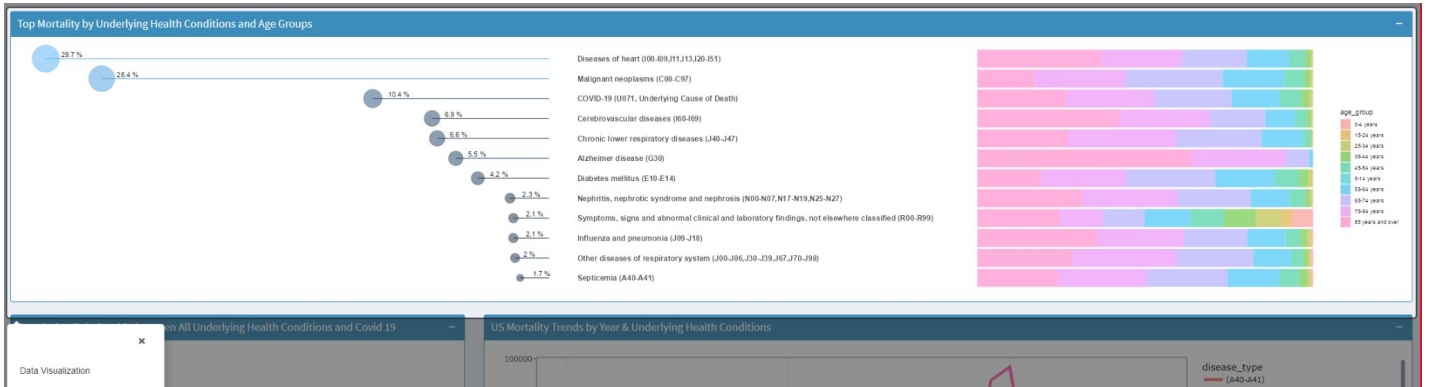
2



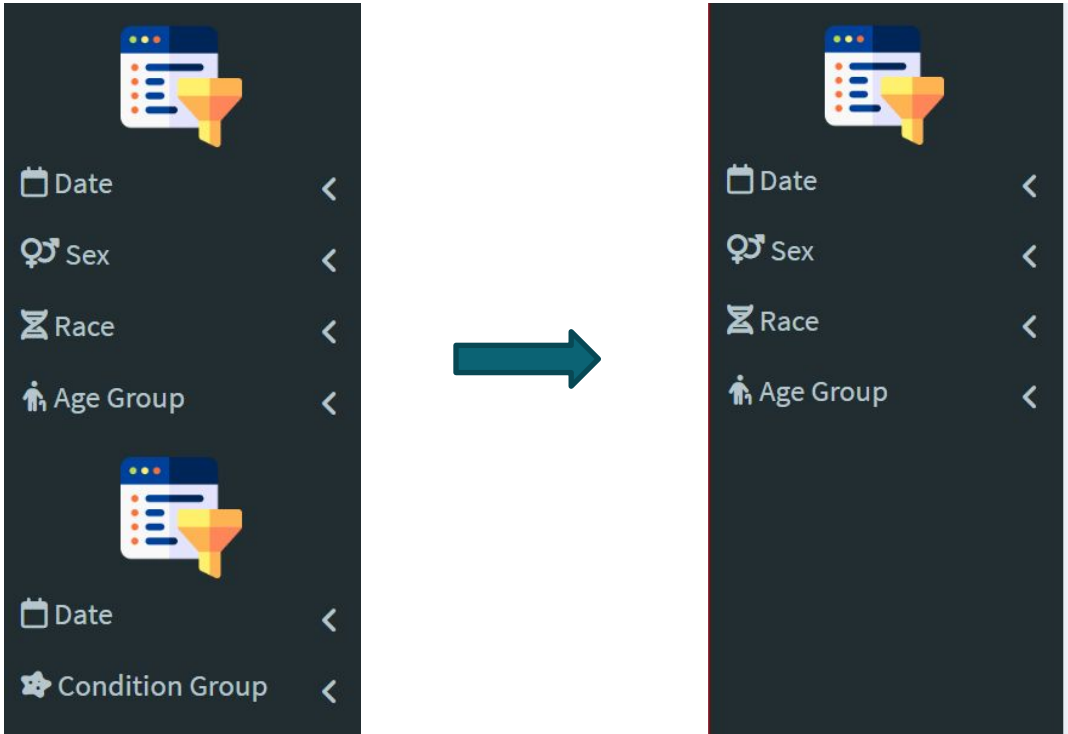
3



4



Shinyjs::show
Shinyjs::hide




```
observeEvent("", {
  shinyjs::show("db1_ui")
  shinyjs::show("db1_panel")
  shinyjs::hide("db2_ui")
  shinyjs::hide("db2_panel")
  shinyjs::hide("db3_ui")
  shinyjs::hide("db3_panel")
  shinyjs::hide("pred_ui")
  shinyjs::hide("pred_panel")
})

observeEvent(input$dataset1, {
  shinyjs::show("db1_ui")
  shinyjs::show("db1_panel")
  shinyjs::hide("db2_ui")
  shinyjs::hide("db2_panel")
  shinyjs::hide("db3_ui")
  shinyjs::hide("db3_panel")
  shinyjs::hide("pred_ui")
  shinyjs::hide("pred_panel")
})

observeEvent(input$dataset2, {
  shinyjs::hide("db1_ui")
  shinyjs::hide("db1_panel")
  shinyjs::show("db2_ui")
  shinyjs::show("db2_panel")
  shinyjs::hide("db3_ui")
  shinyjs::hide("db3_panel")
  shinyjs::hide("pred_ui")
  shinyjs::hide("pred_panel")
})

observeEvent(input$dataset3, {
  shinyjs::hide("db1_ui")
  shinyjs::hide("db1_panel")
  shinyjs::hide("db2_ui")
  shinyjs::hide("db2_panel")
  shinyjs::show("db3_ui")
  shinyjs::show("db3_panel")
  shinyjs::hide("pred_ui")
  shinyjs::hide("pred_panel")
})

observeEvent(input$prediction, {
  shinyjs::hide("db1_ui")
  shinyjs::hide("db1_panel")
  shinyjs::hide("db2_ui")
  shinyjs::hide("db2_panel")
```

ARIMA: Our Dataset

1000 days, from 2020-01-22 to 2022-10-18

```r
summary(usa)
```

```
      date              new_case
 Min.   :2020-01-22   Min.   :      0
 1st Qu.:2020-09-28   1st Qu.:  31122
 Median :2021-06-05   Median :  58628
 Mean   :2021-06-05   Mean   :  96085
 3rd Qu.:2022-02-10   3rd Qu.: 121151
 Max.   :2022-10-18   Max.   :1272910
```

US daily new COVID-19 cases report

```r
usa
```

A tibble: 1,001 × 2

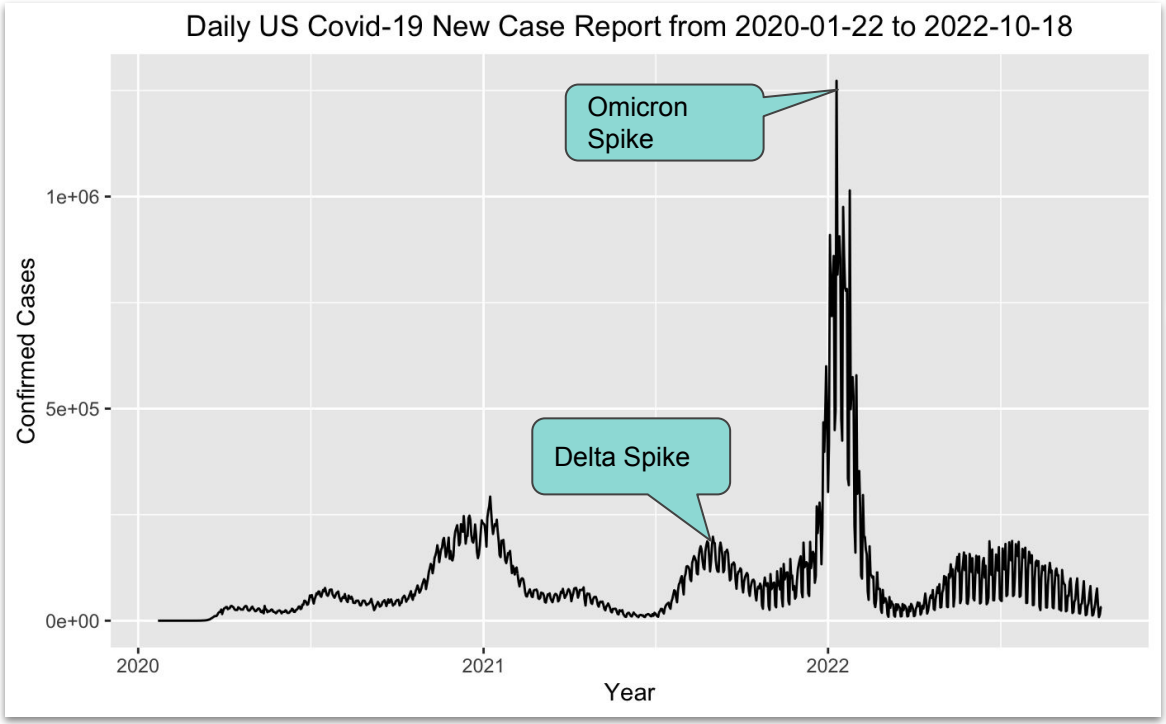
date <date>	new_case <int>
2021-01-16	210088
2021-01-17	187026
2021-01-18	154394
2021-01-19	149895
2021-01-20	188801
2021-01-21	188478
2021-01-22	190409
2021-01-23	166075
2021-01-24	142029
2021-01-25	135794

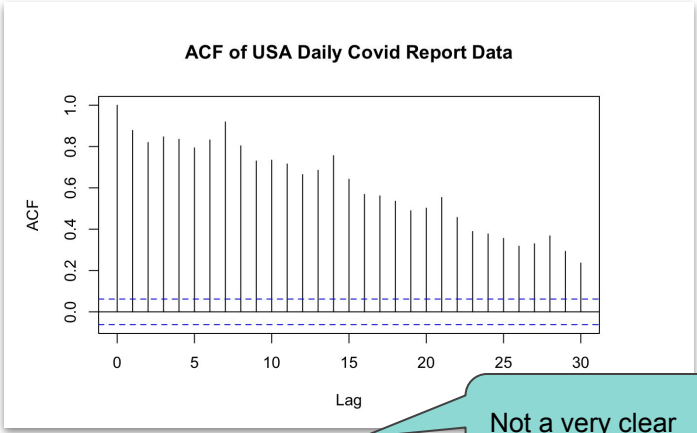
361–370 of 1,001 rows

Sample subset of data

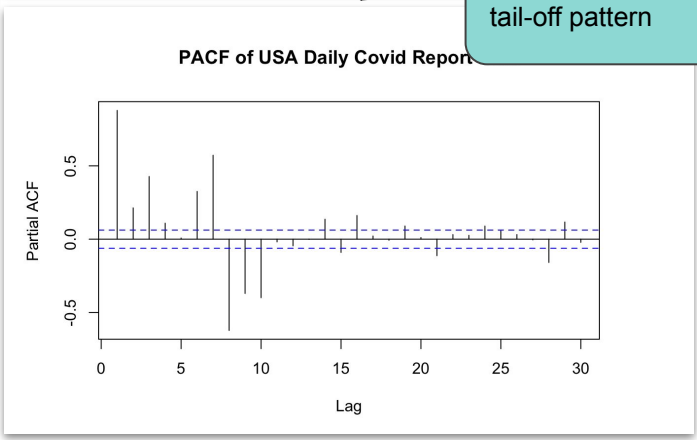
Each day we have a value for new COVID-19 report count

ARIMA: Our Time Series

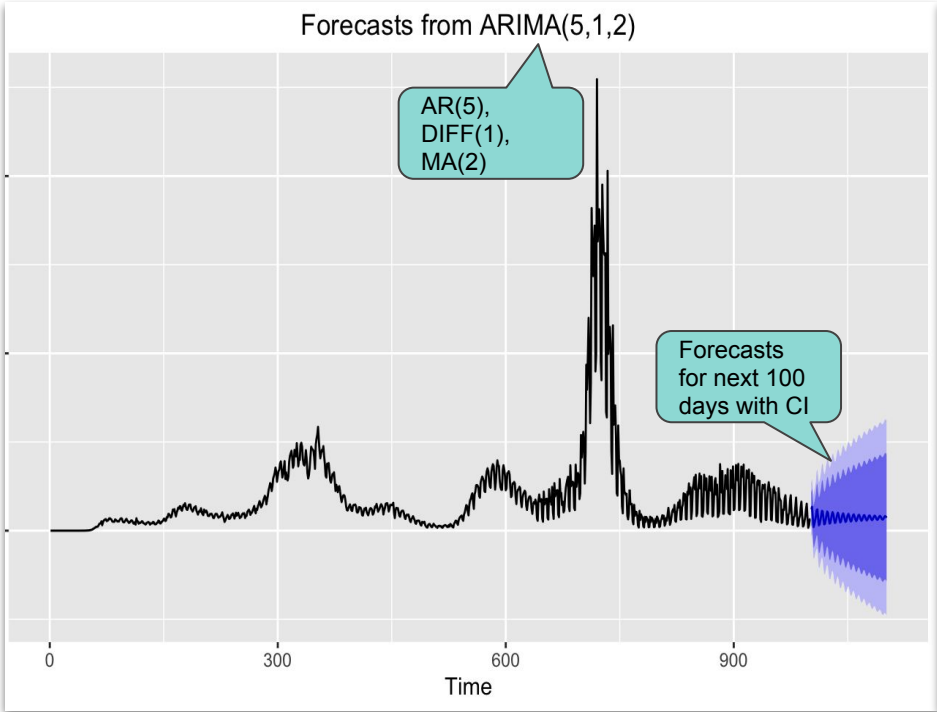


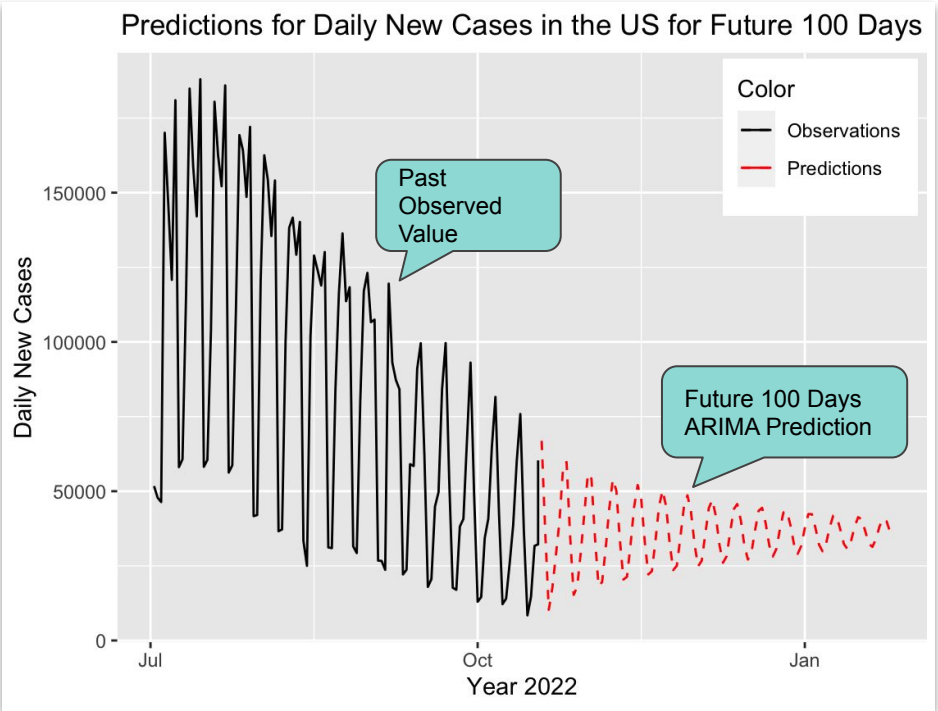


Not a very clear tail-off pattern



ARIMA: Our Model and Forecasts for T+100





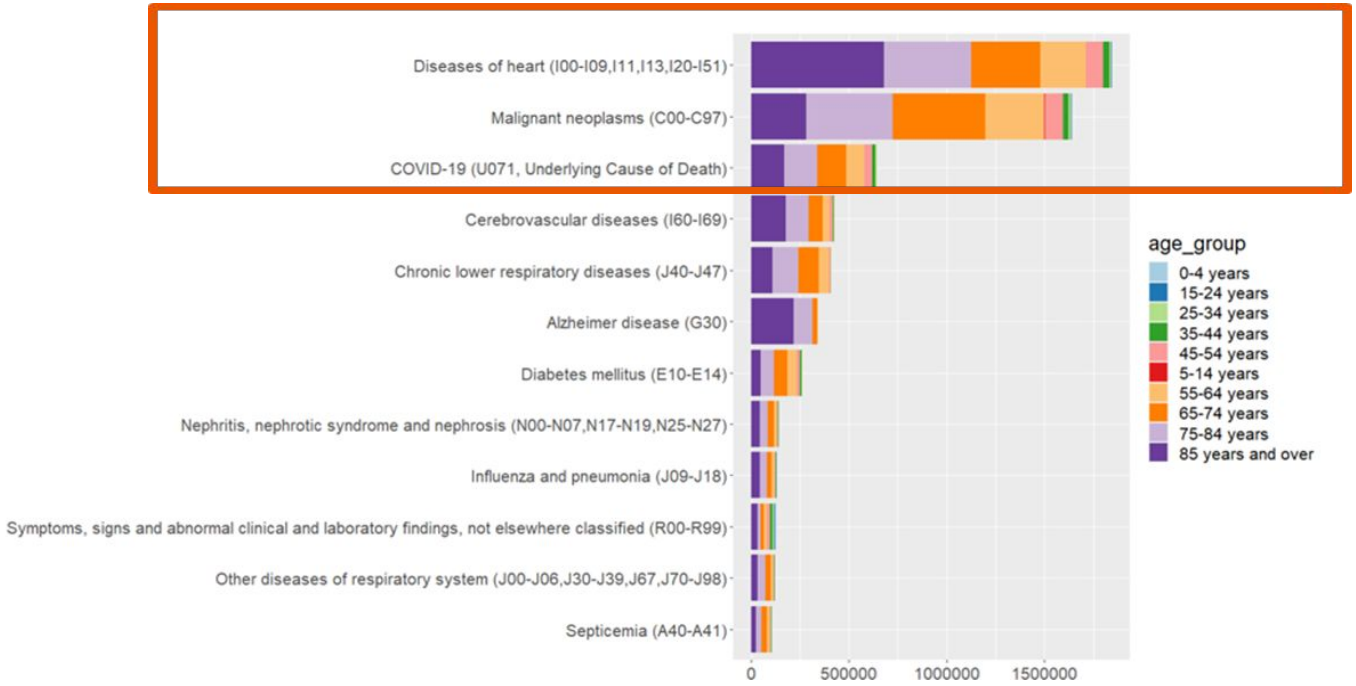
Thank you.

#Now, Appendix

EDA: US Mortality by Top Underlying Health Conditions and Age

Research Question:

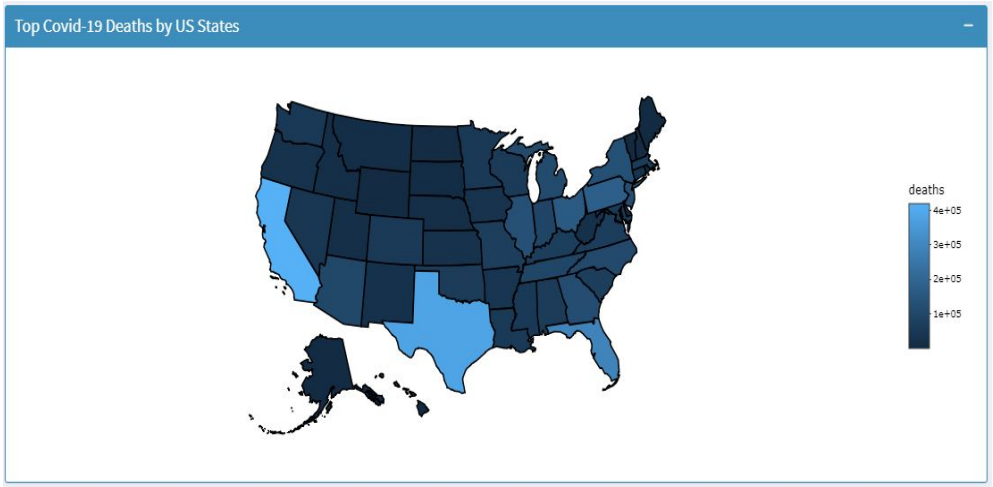
Which underlying medical conditions are most relatable with the national mortality rate in the United States between 2020 – 2022 ?



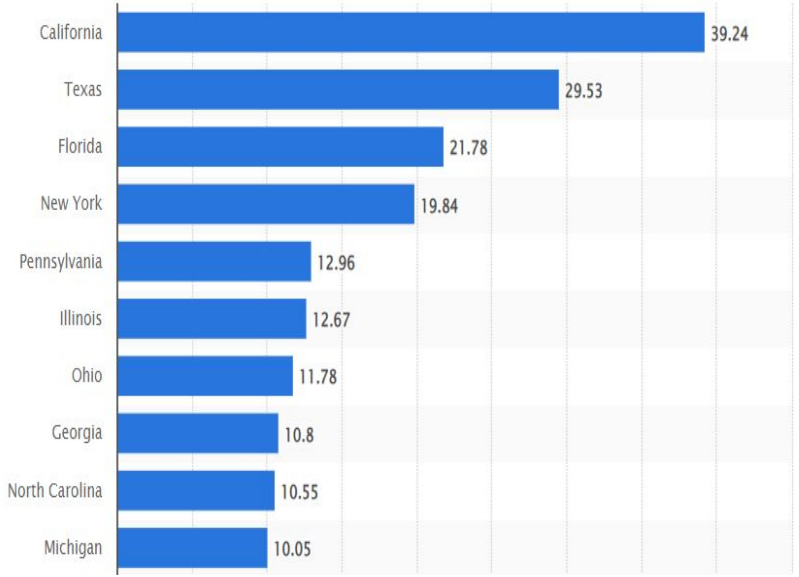
Top three Underlying medical conditions

EDA: Covid – 19 Death Analysis by State

Research Question:
Which states have the highest death rates from Covid-19?



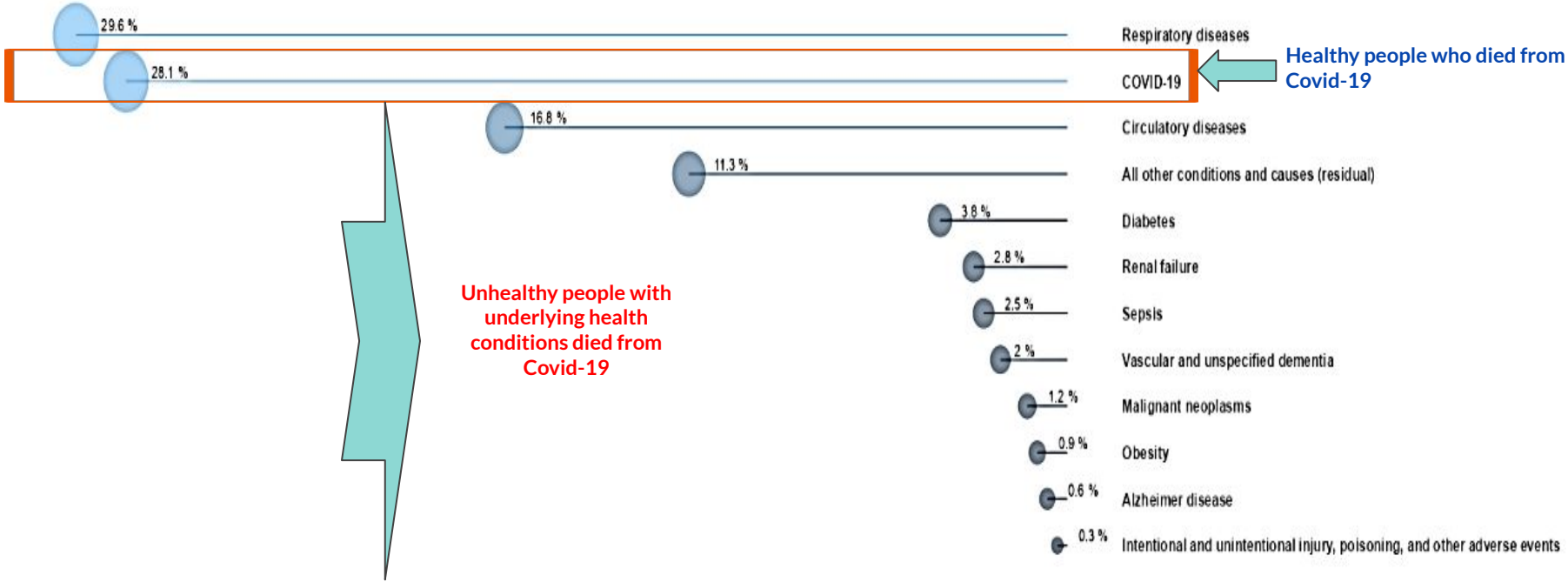
Population of the U.S. in 2021, by State (in millions)



Reference: https://www.statista.com/statistics/183497/population-in-the-federal-states-of-the-us/

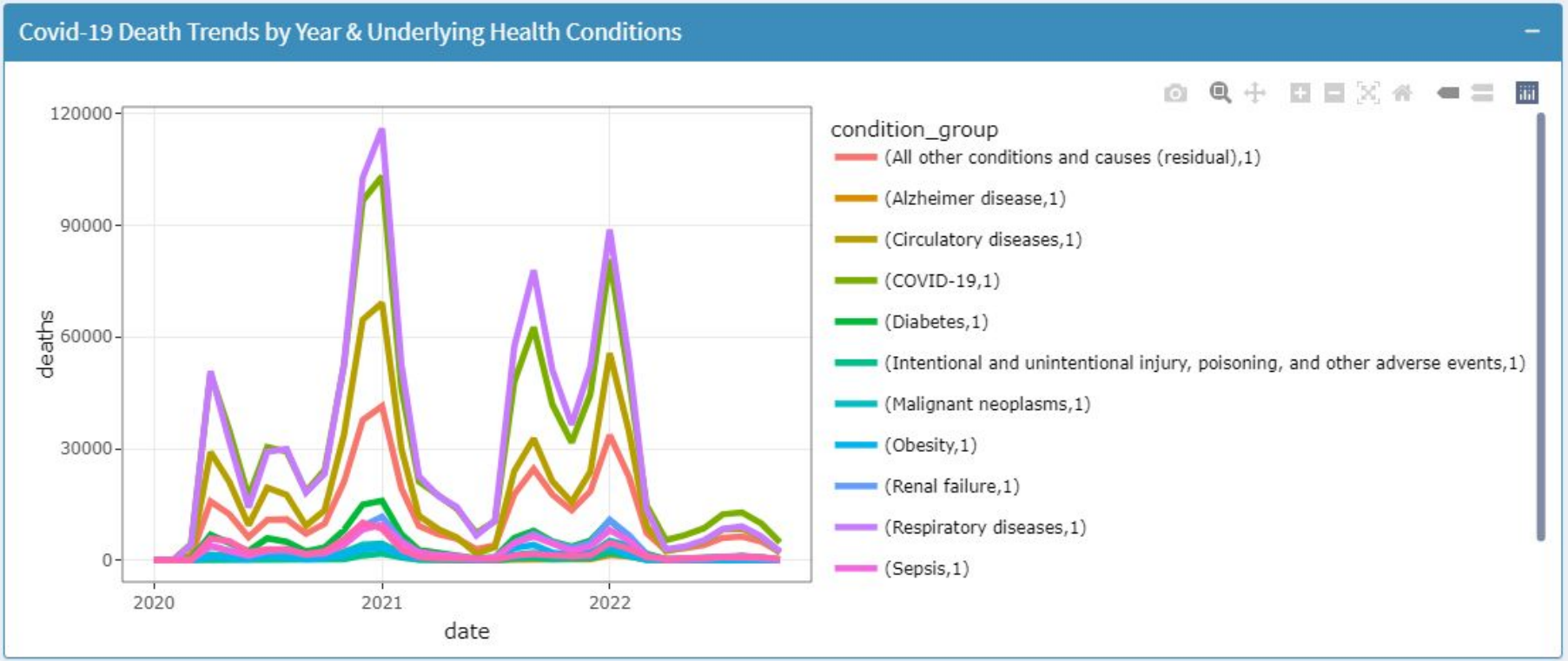
EDA: Covid – 19 Death Analysis by Underlying Health Condition and Age

Research Question:
Who is at high risk from severe coronavirus diseases, and which are top relatable underlying health conditions from Covid-19 ?



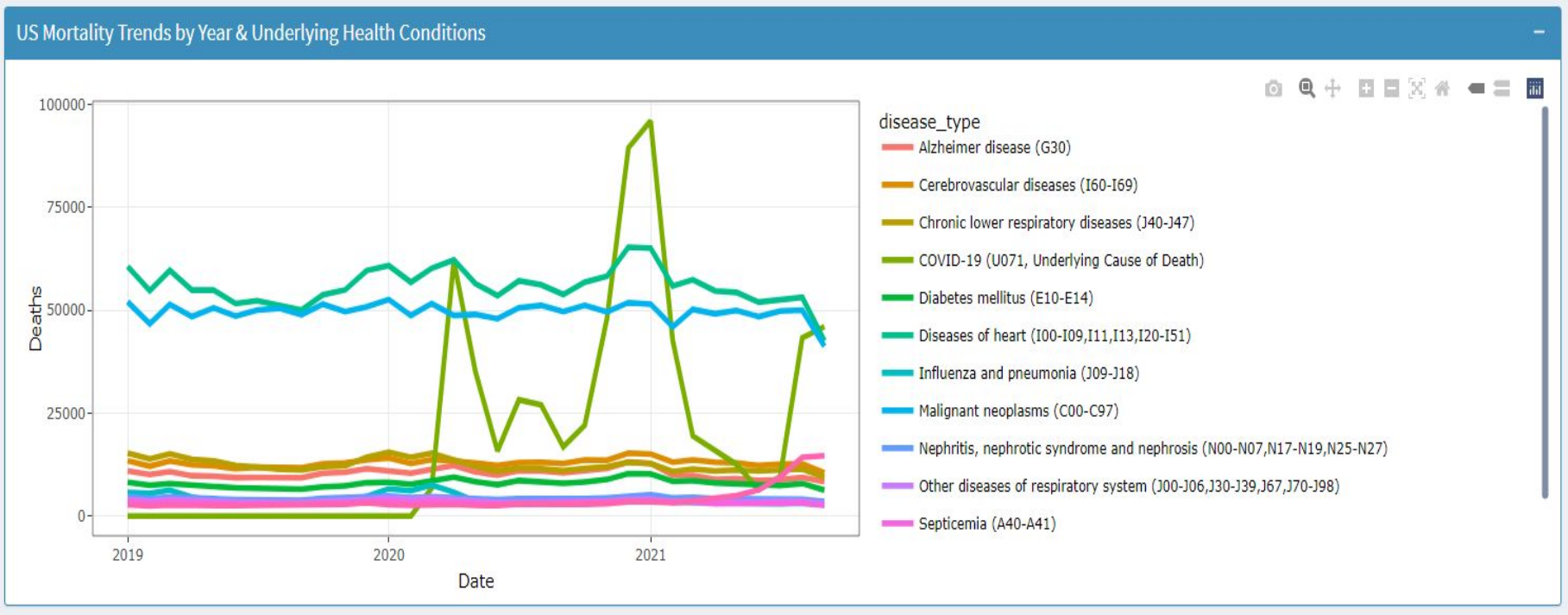
EDA: Covid – 19 Death Analysis by Underlying Health Condition and Age

Research Question: People with which pre-existing health condition are most vulnerable to COVID-19?



EDA: Covid – 19 Death Analysis by Underlying Health Condition and Age

Research Question: Which underlying disease has the most death toll for any given year and what is the correlation among different disease groups?



EDA: Covid – 19 Death Analysis by Underlying Health Condition and Age

Research Question: During which time of year would COVID-19 spread the fastest among people?