Part2_Covid-19_Final_Project_KobyHercsky_Theodore

Theodore Koby-Hercsky

5/22/2021

html_document: https://rpubs.com/theoKoby/772817

Set My working directory set to my final project

setwd("~/Documents/Bellevue University Classes/DSC520/Final Project")

How to import and clean my data

```
## Pulling the Dataframe for the Vaccine administered for Pfizer by state by week data set.

COVID_19_Vaccine_Distribution_Pfizer <- read_csv("CDC Covid Data/COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Pfizer.csv")
View(COVID_19_Vaccine_Distribution_Pfizer)
head(COVID_19_Vaccine_Distribution_Pfizer)

## # A tibble: 6 x 4
##   Jurisdiction  `Week of Allocation… `1st Dose Allocation… `2nd Dose Allocation…
##   <chr>         <chr>                                <dbl>                <dbl>
## 1 Connecticut   05/10/2021                           54990                54990
## 2 Maine         05/10/2021                           21060                21060
## 3 Massachusetts 05/10/2021                          105300               105300
## 4 New Hampshire 05/10/2021                           21060                21060
## 5 Rhode Island  05/10/2021                           16380                16380
## 6 Vermont       05/10/2021                           10530                10530

## Next I will use the str function to look at my data to determine if it is imputed as a character, number, or date.
str(COVID_19_Vaccine_Distribution_Pfizer)

## spec_tbl_df[,4] [1,386 × 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1386] "Connecticut" "Maine" "Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : chr [1:1386] "05/10/2021" "05/10/2021" "05/10/2021" "05/10/2021" ...
```

```
##  $ 1st Dose Allocations: num [1:1386] 54990 21060 105300 21060 16380 ...
##  $ 2nd Dose Allocations: num [1:1386] 54990 21060 105300 21060 16380 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Jurisdiction = col_character(),
##   ..   `Week of Allocations` = col_character(),
##   ..   `1st Dose Allocations` = col_number(),
##   ..   `2nd Dose Allocations` = col_number()
##   .. )
```

```
## As seen above our date is stored as a character which I am going to update to be a date.

COVID_19_Vaccine_Distribution_Pfizer$`Week of Allocations` <- as.Date(COVID_19_Vaccine_Distribution_Pfizer$`Week of Allocations`, format = "%m/%d/%y")
## When we use the str function again we see that the week of allocations is now formatted as a date.
str(COVID_19_Vaccine_Distribution_Pfizer)
```

```
## spec_tbl_df[,4] [1,386 × 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1386] "Connecticut" "Maine" "Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : Date[1:1386], format: "2020-05-10" "2020-05-10" ...
##  $ 1st Dose Allocations: num [1:1386] 54990 21060 105300 21060 16380 ...
##  $ 2nd Dose Allocations: num [1:1386] 54990 21060 105300 21060 16380 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Jurisdiction = col_character(),
##   ..   `Week of Allocations` = col_character(),
##   ..   `1st Dose Allocations` = col_number(),
##   ..   `2nd Dose Allocations` = col_number()
##   .. )
```

```
## After we look over this I will use complete.case and summary function to search for NAs and any other issues we might find with our data.
complete.cases(COVID_19_Vaccine_Distribution_Pfizer)
```

```
##    [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
##   [85] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [99] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [113] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [127] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [141] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [155] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [169] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [183] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [197] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [211] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [225] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [239] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [253] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [267] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [281] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [295] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [309] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [323] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [337] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [351] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [365] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [379] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [393] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [407] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [421] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
```

```
##  [435] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [449] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [463] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [477] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [491] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [505] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [519] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [533] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [547] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [561] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [575] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [589] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [603] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [617] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [631] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [645] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [659] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [673] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [687] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [701] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [715] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [729] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [743] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [757] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [771] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
```

```
##  [785] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [799] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [813] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [827] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [841] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [855] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [869] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [883] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [897] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [911] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [925] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [939] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [953] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [967] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [981] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [995] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1009] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1023] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1037] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1051] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1065] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1079] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1093] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1107] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
```

```
## [1135] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1149] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1163] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1177] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1191] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1205] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1219] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1233] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1247] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1261] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1275] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1289] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1303] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1317] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1331] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1345] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1359] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1373] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE

summary(COVID_19_Vaccine_Distribution_Pfizer)

##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:1386        Min.   :2020-01-04   Min.   :     0
##  Class :character   1st Qu.:2020-02-08   1st Qu.: 11700
##  Mode  :character   Median :2020-03-18   Median : 35100
##                     Mean   :2020-04-15   Mean   : 56990
##                     3rd Qu.:2020-04-26   3rd Qu.: 71906
##                     Max.   :2020-12-28   Max.   :644670
##  2nd Dose Allocations
##  Min.   :     0
##  1st Qu.: 11700
##  Median : 35100
```

```
##  Mean   : 56814
##  3rd Qu.: 71906
##  Max.   :644670
```

```
## I notice that we have values that are zero for 1st dose allocation which
we are going to remove due to the fact that these values are not justifiable
as the can be bad data.

COVID_19_Vaccine_Distribution_Pfizer <-
subset(COVID_19_Vaccine_Distribution_Pfizer,
COVID_19_Vaccine_Distribution_Pfizer$`1st Dose Allocations` >= "1")

## Next I will be removing the variable 2nd Dose Allocations as these values
are the same as the values in the 1st Dose Allocations
COVID_19_Vaccine_Distribution_Pfizer$`2nd Dose Allocations` <- NULL

## I also want to add a new variable to my dataframe that is simply the name
of the companies shot.
COVID_19_Vaccine_Distribution_Pfizer$Company <- "Pfizer"

## After all the necessary changes have been made we can now use the str and
summary functions to verify our data before moving on to our next data frame.
str(COVID_19_Vaccine_Distribution_Pfizer)
```

```
## tibble[,4] [1,254 × 4] (S3: tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1254] "Connecticut" "Maine"
"Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : Date[1:1254], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:1254] 54990 21060 105300 21060 16380 ...
##  $ Company             : chr [1:1254] "Pfizer" "Pfizer" "Pfizer" "Pfizer"
...
```

```
summary(COVID_19_Vaccine_Distribution_Pfizer)
```

```
##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:1254        Min.   :2020-01-04   Min.   :   975
##  Class :character   1st Qu.:2020-02-08   1st Qu.: 17550
##  Mode  :character   Median :2020-03-15   Median : 39780
##                     Mean   :2020-04-15   Mean   : 62989
##                     3rd Qu.:2020-04-26   3rd Qu.: 77171
##                     Max.   :2020-12-28   Max.   :644670
##    Company
##  Length:1254
##  Class :character
##  Mode  :character
##
##
##
```

```
## Also when viewing this data frame I can filter the 1st dose to determine
which states are giving the most doses of the Pfizer vaccine as it showed
California was the highest


## Pulling the Dataframe for the Vaccine administered for Moderna by state by
week data set

COVID_19_Vaccine_Distribution_Moderna <- read_csv("CDC Covid Data/COVID-
19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Moderna.csv")
View(COVID_19_Vaccine_Distribution_Moderna)

## Next I will use the str function to look at my data to determine if it is
imputed as a character, number, or date.
str(COVID_19_Vaccine_Distribution_Moderna)

## spec_tbl_df[,4] [1,323 × 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1323] "Connecticut" "Maine"
"Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : chr [1:1323] "05/10/2021" "05/10/2021"
"05/10/2021" "05/10/2021" ...
##  $ 1st Dose Allocations: num [1:1323] 41300 15800 79500 15900 12400 ...
##  $ 2nd Dose Allocations: num [1:1323] 41300 15800 79500 15900 12400 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Jurisdiction = col_character(),
##   ..   `Week of Allocations` = col_character(),
##   ..   `1st Dose Allocations` = col_number(),
##   ..   `2nd Dose Allocations` = col_number()
##   .. )

## As seen above our date is stored as a character which I am going to update
to be a date.

COVID_19_Vaccine_Distribution_Moderna$`Week of Allocations` <-
as.Date(COVID_19_Vaccine_Distribution_Moderna$`Week of Allocations`, format =
"%m/%d/%y")
## When we use the str function again we see that the week of allocations is
now formatted as a date.
str(COVID_19_Vaccine_Distribution_Moderna)

## spec_tbl_df[,4] [1,323 × 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1323] "Connecticut" "Maine"
"Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : Date[1:1323], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:1323] 41300 15800 79500 15900 12400 ...
##  $ 2nd Dose Allocations: num [1:1323] 41300 15800 79500 15900 12400 ...
##  - attr(*, "spec")=
##   .. cols(
```

```
##   ..   Jurisdiction = col_character(),
##   ..   `Week of Allocations` = col_character(),
##   ..   `1st Dose Allocations` = col_number(),
##   ..   `2nd Dose Allocations` = col_number()
##   .. )

## After we look over this I will use complete.case and summary function to search for NAs and any other issues we might find with our data.
complete.cases(COVID_19_Vaccine_Distribution_Moderna)

##    [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [85] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##   [99] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [113] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [127] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [141] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [155] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [169] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [183] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [197] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [211] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [225] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [239] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [253] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [267] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
```

```
##  [281] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [295] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [309] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [323] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [337] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [351] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [365] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [379] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [393] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [407] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [421] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [435] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [449] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [463] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [477] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [491] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [505] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [519] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [533] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [547] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [561] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [575] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [589] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [603] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [617] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
```

```
##  [631] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [645] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [659] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [673] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [687] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [701] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [715] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [729] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [743] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [757] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [771] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [785] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [799] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [813] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [827] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [841] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [855] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [869] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [883] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [897] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [911] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [925] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [939] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [953] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [967] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
```

```
##  [981] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
##  [995] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1009] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1023] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1037] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1051] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1065] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1079] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1093] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1107] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1135] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1149] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1163] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1177] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1191] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1205] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1219] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1233] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1247] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1261] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1275] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1289] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1303] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [1317] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
summary(COVID_19_Vaccine_Distribution_Moderna)
```

```
##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:1323        Min.   :2020-01-04   Min.   :     0
##  Class :character   1st Qu.:2020-02-08   1st Qu.: 11100
##  Mode  :character   Median :2020-03-15   Median : 31600
##                     Mean   :2020-04-04   Mean   : 51822
##                     3rd Qu.:2020-04-19   3rd Qu.: 66700
##                     Max.   :2020-12-28   Max.   :672600
##  2nd Dose Allocations
##  Min.   :     0
##  1st Qu.: 10800
##  Median : 31600
##  Mean   : 51621
##  3rd Qu.: 66700
##  Max.   :672600
```

```
## I notice that we have values that are zero for 1st dose allocation which we are going to remove due to the fact that these values are not justifiable as the can be bad data.

COVID_19_Vaccine_Distribution_Moderna <-
subset(COVID_19_Vaccine_Distribution_Moderna,
COVID_19_Vaccine_Distribution_Moderna$`1st Dose Allocations` >= "1")

## Next I will be removing the variable 2nd Dose Allocations as these values are the same as the values in the 1st Dose Allocations
COVID_19_Vaccine_Distribution_Moderna$`2nd Dose Allocations` <- NULL

## I also want to add a new variable to my dataframe that is simply the name of the companies shot.
COVID_19_Vaccine_Distribution_Moderna$Company <- "Moderna"

## After all the necessary changes have been made we can now use the str and summary functions to verify our data before moving on to our next data frame.
str(COVID_19_Vaccine_Distribution_Moderna)
```

```
## tibble[,4] [1,213 × 4] (S3: tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1213] "Connecticut" "Maine"
"Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : Date[1:1213], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:1213] 41300 15800 79500 15900 12400 ...
##  $ Company             : chr [1:1213] "Moderna" "Moderna" "Moderna"
"Moderna" ...
```

```
summary(COVID_19_Vaccine_Distribution_Moderna)
```

```
##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:1213        Min.   :2020-01-04   Min.   :   700
##  Class :character   1st Qu.:2020-02-08   1st Qu.: 15600
```

```
##  Mode  :character   Median :2020-03-15   Median : 35800
##                     Mean   :2020-04-04   Mean   : 56522
##                     3rd Qu.:2020-04-19   3rd Qu.: 69000
##                     Max.   :2020-12-28   Max.   :672600
##    Company
##  Length:1213
##  Class :character
##  Mode  :character
##
##
##
```

```
## Pulling the Dataframe for the Vaccine administered for Johnson and Johnson by state by week data set.
COVID_19_Vaccine_Distribution_Johnson <- read_csv("CDC Covid Data/COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Janssen.csv")
View(COVID_19_Vaccine_Distribution_Johnson)
head(COVID_19_Vaccine_Distribution_Johnson)
```

```
## # A tibble: 6 x 3
##   Jurisdiction  `Week of Allocations` `1st Dose Allocations`
##   <chr>         <chr>                                 <dbl>
## 1 Connecticut   05/10/2021                             6400
## 2 Maine         05/10/2021                             2500
## 3 Massachusetts 05/10/2021                            12300
## 4 New Hampshire 05/10/2021                             2500
## 5 Rhode Island  05/10/2021                             2000
## 6 Vermont       05/10/2021                             1200
```

```
## Next I will use the str function to look at my data to determine if it is imputed as a character, number, or date.
str(COVID_19_Vaccine_Distribution_Johnson)
```

```
## spec_tbl_df[,3] [504 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:504] "Connecticut" "Maine" "Massachusetts"
"New Hampshire" ...
##  $ Week of Allocations : chr [1:504] "05/10/2021" "05/10/2021"
"05/10/2021" "05/10/2021" ...
##  $ 1st Dose Allocations: num [1:504] 6400 2500 12300 2500 2000 1200 15600
19800 15100 6100 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Jurisdiction = col_character(),
##   ..   `Week of Allocations` = col_character(),
##   ..   `1st Dose Allocations` = col_number()
##   .. )
```

```
## As seen above our date is stored as a character which I am going to update to be a date.

COVID_19_Vaccine_Distribution_Johnson$`Week of Allocations` <-
```

```
as.Date(COVID_19_Vaccine_Distribution_Johnson$`Week of Allocations`, format =
"%m/%d/%y")
## When we use the str function again we see that the week of allocations is
now formatted as a date.
str(COVID_19_Vaccine_Distribution_Johnson)

## spec_tbl_df[,3] [504 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:504] "Connecticut" "Maine" "Massachusetts"
"New Hampshire" ...
##  $ Week of Allocations : Date[1:504], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:504] 6400 2500 12300 2500 2000 1200 15600
19800 15100 6100 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Jurisdiction = col_character(),
##   ..   `Week of Allocations` = col_character(),
##   ..   `1st Dose Allocations` = col_number()
##   .. )

## After we look over this I will use complete.case and summary function to
search for NAs and any other issues we might find with our data.
complete.cases(COVID_19_Vaccine_Distribution_Johnson)

##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [136] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [151] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [166] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [181] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
```

```
## [196] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [211] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [226] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [241] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [256] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [271] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [286] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [301] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [316] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [331] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [346] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [361] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [376] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [391] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [406] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [421] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [436] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [451] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [466] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [481] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [496] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

summary(COVID_19_Vaccine_Distribution_Johnson)

##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:504         Min.   :2020-03-01   Min.   :     0
##  Class :character   1st Qu.:2020-03-20   1st Qu.:  3200
##  Mode  :character   Median :2020-04-01   Median :  9500
##                     Mean   :2020-04-04   Mean   : 25089
```

```
##                     3rd Qu.:2020-04-17   3rd Qu.: 24100
##                     Max.   :2020-05-10   Max.   :572700

## I notice that we have values that are zero for 1st dose allocation which
we are going to remove due to the fact that these values are not justifiable
as the can be bad data.

COVID_19_Vaccine_Distribution_Johnson <-
subset(COVID_19_Vaccine_Distribution_Johnson,
COVID_19_Vaccine_Distribution_Johnson$`1st Dose Allocations` >= "1")

## I also want to add a new variable to my dataframe that is simply the name
of the companies shot.
COVID_19_Vaccine_Distribution_Johnson$Company <- "Johnson"

## After all the necessary changes have been made we can now use the str and
summary functions to verify our data before moving on to our next data frame.
str(COVID_19_Vaccine_Distribution_Johnson)

## tibble[,4] [486 × 4] (S3: tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:486] "Connecticut" "Maine" "Massachusetts"
"New Hampshire" ...
##  $ Week of Allocations : Date[1:486], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:486] 6400 2500 12300 2500 2000 1200 15600
19800 15100 6100 ...
##  $ Company             : chr [1:486] "Johnson" "Johnson" "Johnson"
"Johnson" ...

summary(COVID_19_Vaccine_Distribution_Johnson)

##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:486         Min.   :2020-03-01   Min.   :   100
##  Class :character   1st Qu.:2020-03-15   1st Qu.:  3400
##  Mode  :character   Median :2020-03-29   Median :  9750
##                     Mean   :2020-04-03   Mean   : 26018
##                     3rd Qu.:2020-04-12   3rd Qu.: 25300
##                     Max.   :2020-05-10   Max.   :572700
##    Company
##  Length:486
##  Class :character
##  Mode  :character
##
##
##

## The first Data Set is the Covid-19 Deaths by race, sex, and age data frame
that I will pull and clean
COVID_19_Deaths_by_Educational_Race_Sex_and_Age <- read_csv("CDC Covid
Data/AH_Provisional_COVID-
19_Deaths_by_Educational_Attainment__Race__Sex__and_Age.csv")
```

```
View(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
head(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
```

```
## # A tibble: 6 x 9
##   `Data as of` `Start Date` `End Date` `Education Level`  `Race or Hispan…
Sex
##   <chr>        <chr>        <chr>      <chr>              <chr>
<chr>
## 1 02/03/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 2 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 3 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 4 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 5 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Male
## 6 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Male
## # … with 3 more variables: Age Group <chr>, COVID-19 Deaths <dbl>,
## #   Total Deaths <dbl>
```

```
## After viewing this I am going to delete the variable date as of and also
use the names function to update the variables names that are needed.
## Delete variable date as of.
COVID_19_Deaths_by_Educational_Race_Sex_and_Age$`Data as of` <- NULL
## Change variable names
names(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)[3]<-"Education"
names(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)[4]<-"Race"

## Next I will use the str function to look at my data to determine if it is
imputed as a character, number, or date.
str(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
```

```
## spec_tbl_df[,8] [224 × 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Start Date     : chr [1:224] "01/01/2020" "01/01/2020" "01/01/2020"
"01/01/2020" ...
##  $ End Date       : chr [1:224] "01/30/2021" "01/30/2021" "01/30/2021"
"01/30/2021" ...
##  $ Education      : chr [1:224] "Associate degree or some college"
"Associate degree or some college" "Associate degree or some college"
"Associate degree or some college" ...
##  $ Race           : chr [1:224] "Hispanic" "Hispanic" "Hispanic"
"Hispanic" ...
##  $ Sex            : chr [1:224] "Female" "Female" "Female" "Female" ...
##  $ Age Group      : chr [1:224] "0-17 years" "18-49 years" "50-64 years"
"65 years and over" ...
##  $ COVID-19 Deaths: num [1:224] 0 423 857 1793 0 ...
##  $ Total Deaths   : num [1:224] 2 3117 4153 10225 1 ...
```

```
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Data as of` = col_character(),
##   ..   `Start Date` = col_character(),
##   ..   `End Date` = col_character(),
##   ..   `Education Level` = col_character(),
##   ..   `Race or Hispanic Origin` = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age Group` = col_character(),
##   ..   `COVID-19 Deaths` = col_number(),
##   ..   `Total Deaths` = col_number()
##   .. )
```

```
## As seen above our dates are stored as a character which I am going to
update to be a date and also delete the date as of as that is unneeded.

## change the format of the start date
COVID_19_Deaths_by_Educational_Race_Sex_and_Age$`Start Date` <-
as.Date(COVID_19_Deaths_by_Educational_Race_Sex_and_Age$`Start Date`, format
= "%m/%d/%y")
## Change the format of the end date
COVID_19_Deaths_by_Educational_Race_Sex_and_Age$`End Date` <-
as.Date(COVID_19_Deaths_by_Educational_Race_Sex_and_Age$`End Date`, format =
"%m/%d/%y")

## I also noticed earlier that the variable education has some Unknown values
which I would like to remove due to the fact that these values are not
justifiable as the can be bad data.

## As seen below I removed any value that was Unknown in Education
COVID_19_Deaths_by_Educational_Race_Sex_and_Age <-
COVID_19_Deaths_by_Educational_Race_Sex_and_Age[COVID_19_Deaths_by_Educationa
l_Race_Sex_and_Age$Education !="Unknown", ]

## After we look over this I will use complete.case and summary function to
search for NAs and any other issues we might find with our data.
complete.cases(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
```

```
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [136] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [151] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [166] TRUE TRUE TRUE

summary(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)

##    Start Date            End Date           Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:168
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##      Race               Sex             Age Group         COVID-19 Deaths
##  Length:168         Length:168         Length:168         Min.   :    0.0
##  Class :character   Class :character   Class :character   1st Qu.:   12.5
##  Mode  :character   Mode  :character   Mode  :character   Median :  133.0
##                                                           Mean   : 2441.9
##                                                           3rd Qu.:  958.5
##                                                           Max.   :76871.0
##   Total Deaths
##  Min.   :     0.0
##  1st Qu.:   159.2
##  Median :  1255.0
##  Mean   : 20479.5
##  3rd Qu.:  8724.8
##  Max.   :670295.0

## Next I will pull the data frame Covid-19 Deaths by education and race and
filter and clean up this data.
COVID_19_Deaths_by_Race_and_Educational <- read_csv("CDC Covid
Data/AH_Provisional_COVID-19_Deaths_by_Race_and_Educational_Attainment.csv")
View(COVID_19_Deaths_by_Race_and_Educational)
head(COVID_19_Deaths_by_Race_and_Educational)

## # A tibble: 6 x 7
##   `Data as of` `Start Date` `End Date` `Education Level` `Race or Hispanic
Orig…
##   <chr>        <chr>        <chr>      <chr>             <chr>
## 1 02/01/2021   01/01/2020   01/30/2021 8th grade or less Hispanic
## 2 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
American …
## 3 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
```

```
Asian
## 4 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
Black
## 5 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic More
than…
## 6 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
Native Ha…
## # … with 2 more variables: COVID-19 Deaths <dbl>, Total Deaths <dbl>

## After viewing this I am going to Delete variable date as of and use the
names function to update the variables names that are needed.
COVID_19_Deaths_by_Race_and_Educational$`Data as of` <- NULL
names(COVID_19_Deaths_by_Race_and_Educational)[3]<-"Education"
names(COVID_19_Deaths_by_Race_and_Educational)[4]<-"Race"

## Next I will use the str function to look at my data to determine if it is
imputed as a character, number, or date.
str(COVID_19_Deaths_by_Race_and_Educational)

## spec_tbl_df[,6] [72 × 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Start Date     : chr [1:72] "01/01/2020" "01/01/2020" "01/01/2020"
"01/01/2020" ...
##  $ End Date       : chr [1:72] "01/30/2021" "01/30/2021" "01/30/2021"
"01/30/2021" ...
##  $ Education      : chr [1:72] "8th grade or less" "8th grade or less"
"8th grade or less" "8th grade or less" ...
##  $ Race           : chr [1:72] "Hispanic" "Non-Hispanic American Indian or
Alaska Native" "Non-Hispanic Asian" "Non-Hispanic Black" ...
##  $ COVID-19 Deaths: num [1:72] 29157 706 2610 5699 103 ...
##  $ Total Deaths   : num [1:72] 106285 3085 16283 41437 1676 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Data as of` = col_character(),
##   ..   `Start Date` = col_character(),
##   ..   `End Date` = col_character(),
##   ..   `Education Level` = col_character(),
##   ..   `Race or Hispanic Origin` = col_character(),
##   ..   `COVID-19 Deaths` = col_number(),
##   ..   `Total Deaths` = col_number()
##   .. )

## As seen above our dates are stored as a character which I am going to
update to be a date.

## change the format of the start date
COVID_19_Deaths_by_Race_and_Educational$`Start Date` <-
as.Date(COVID_19_Deaths_by_Race_and_Educational$`Start Date`, format =
"%m/%d/%y")
## Change the format of the end date
COVID_19_Deaths_by_Race_and_Educational$`End Date` <-
```

```
as.Date(COVID_19_Deaths_by_Race_and_Educational$`End Date`, format =
"%m/%d/%y")

## As seen below I removed any value that was Unknown in Education
COVID_19_Deaths_by_Race_and_Educational <-
COVID_19_Deaths_by_Race_and_Educational[COVID_19_Deaths_by_Race_and_Education
al$Education !="Unknown", ]

## The final step we will take for this data frame is to use complete.case
and summary function to double check our data before moving on.
complete.cases(COVID_19_Deaths_by_Race_and_Educational)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE
## [61] TRUE TRUE TRUE TRUE
```

```
summary(COVID_19_Deaths_by_Race_and_Educational)
```

```
##    Start Date            End Date            Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:64
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##      Race           COVID-19 Deaths      Total Deaths
##  Length:64          Min.   :     3.00   Min.   :     21.0
##  Class :character   1st Qu.:    89.25   1st Qu.:    567.2
##  Mode  :character   Median :   684.00   Median :   3907.5
##                     Mean   :  6287.28   Mean   :  53335.8
##                     3rd Qu.:  4577.50   3rd Qu.:  29872.5
##                     Max.   :117989.00   Max.   :1145594.0
```

```
## The final data set I will be pulling and cleaning is the Covid-19 Deaths
by age in years

COVID_19_Deaths_Counts_by_Age_in_Years <- read_csv("CDC Covid
Data/Provisional_COVID-19_Deaths_Counts_by_Age_in_Years.csv")
View(COVID_19_Deaths_Counts_by_Age_in_Years)
head(COVID_19_Deaths_Counts_by_Age_in_Years)
```

```
## # A tibble: 6 x 7
##   `Data as of` `Start Date` `End Date` Sex   `Age Years` `Total deaths`
##   <chr>        <chr>        <chr>      <chr> <chr>                <dbl>
## 1 05/10/2021   01/01/2020   04/30/2021 Male  <1 year              13525
```

```
## 2 05/10/2021   01/01/2020   04/30/2021 Male  01 Years                949
## 3 05/10/2021   01/01/2020   04/30/2021 Male  02 Years                635
## 4 05/10/2021   01/01/2020   04/30/2021 Male  03 Years                515
## 5 05/10/2021   01/01/2020   04/30/2021 Male  04 Years                406
## 6 05/10/2021   01/01/2020   04/30/2021 Male  05 Years                381
## # … with 1 more variable: COVID-19 Deaths <dbl>
```

```
## First I will be deleting the variable date as of.
COVID_19_Deaths_Counts_by_Age_in_Years$`Data as of` <- NULL
```

```
## Next I will use the str function to look at my data to determine if it is
imputed as a character, number, or date.
str(COVID_19_Deaths_Counts_by_Age_in_Years)
```

```
## spec_tbl_df[,6] [172 × 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Start Date     : chr [1:172] "01/01/2020" "01/01/2020" "01/01/2020"
"01/01/2020" ...
##  $ End Date       : chr [1:172] "04/30/2021" "04/30/2021" "04/30/2021"
"04/30/2021" ...
##  $ Sex            : chr [1:172] "Male" "Male" "Male" "Male" ...
##  $ Age Years      : chr [1:172] "<1 year" "01 Years" "02 Years" "03 Years"
...
##  $ Total deaths   : num [1:172] 13525 949 635 515 406 ...
##  $ COVID-19 Deaths: num [1:172] 40 11 3 3 4 6 1 4 5 6 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Data as of` = col_character(),
##   ..   `Start Date` = col_character(),
##   ..   `End Date` = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age Years` = col_character(),
##   ..   `Total deaths` = col_number(),
##   ..   `COVID-19 Deaths` = col_number()
##   .. )
```

```
## As seen above our dates are stored as a character which I am going to
update to be a date.

## change the format of the start date
COVID_19_Deaths_Counts_by_Age_in_Years$`Start Date` <-
as.Date(COVID_19_Deaths_Counts_by_Age_in_Years$`Start Date`, format =
"%m/%d/%y")
## Change the format of the end date
COVID_19_Deaths_Counts_by_Age_in_Years$`End Date` <-
as.Date(COVID_19_Deaths_Counts_by_Age_in_Years$`End Date`, format =
"%m/%d/%y")

## I am going to create a new variable to add new insights on the percentage
of deaths that are from Covid-19.
```

```
COVID_19_Deaths_Counts_by_Age_in_Years["COVID-19 Death Percentage"] <-
COVID_19_Deaths_Counts_by_Age_in_Years$`COVID-19
Deaths`/COVID_19_Deaths_Counts_by_Age_in_Years$`Total deaths`

## The final step we will take for this data frame is to use complete.case
and summary function to double check our data before moving on.
complete.cases(COVID_19_Deaths_Counts_by_Age_in_Years)

##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [136] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [151] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
TRUE TRUE
## [166] TRUE TRUE TRUE TRUE TRUE TRUE TRUE

summary(COVID_19_Deaths_Counts_by_Age_in_Years)

##    Start Date            End Date              Sex
##  Min.   :2020-01-01   Min.   :2020-04-30   Length:172
##  1st Qu.:2020-01-01   1st Qu.:2020-04-30   Class :character
##  Median :2020-01-01   Median :2020-04-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-04-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-04-30
##  Max.   :2020-01-01   Max.   :2020-04-30
##   Age Years          Total deaths    COVID-19 Deaths
##  Length:172         Min.   :   214   Min.   :    1.00
##  Class :character   1st Qu.:  2296   1st Qu.:   55.25
##  Mode  :character   Median :  8826   Median :  598.00
##                     Mean   : 25897   Mean   : 3289.17
##                     3rd Qu.: 34984   3rd Qu.: 4449.50
##                     Max.   :802843   Max.   :98328.00
##  COVID-19 Death Percentage
##  Min.   :0.002141
```

```
##  1st Qu.:0.023516
##  Median :0.085025
##  Mean   :0.078166
##  3rd Qu.:0.126495
##  Max.   :0.158603
```

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

What does the final data set look like?

With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

```
## I determined the best way to show the most condensed version of my data is
to use the str() function which shows the format of the data and a couple
values from each variable. While indicating if the variable is a date,
character, and or number.

## The first str is for the data frame for the Pfizer vaccine
str(COVID_19_Vaccine_Distribution_Pfizer)

## tibble[,4] [1,254 × 4] (S3: tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1254] "Connecticut" "Maine"
"Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : Date[1:1254], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:1254] 54990 21060 105300 21060 16380 ...
##  $ Company             : chr [1:1254] "Pfizer" "Pfizer" "Pfizer" "Pfizer"
...

## The next str is for the data frame for the Moderna vaccine
str(COVID_19_Vaccine_Distribution_Moderna)

## tibble[,4] [1,213 × 4] (S3: tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:1213] "Connecticut" "Maine"
"Massachusetts" "New Hampshire" ...
##  $ Week of Allocations : Date[1:1213], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:1213] 41300 15800 79500 15900 12400 ...
##  $ Company             : chr [1:1213] "Moderna" "Moderna" "Moderna"
"Moderna" ...

## The following str is for the data frame for the Johnson & Johnson vaccine
str(COVID_19_Vaccine_Distribution_Johnson)

## tibble[,4] [486 × 4] (S3: tbl_df/tbl/data.frame)
##  $ Jurisdiction        : chr [1:486] "Connecticut" "Maine" "Massachusetts"
"New Hampshire" ...
##  $ Week of Allocations : Date[1:486], format: "2020-05-10" "2020-05-10"
...
##  $ 1st Dose Allocations: num [1:486] 6400 2500 12300 2500 2000 1200 15600
```

```
19800 15100 6100 ...
##  $ Company             : chr [1:486] "Johnson" "Johnson" "Johnson"
"Johnson" ...

## This str is for the data frame for the COVID-19 Death rate that is
organized by the education, race, sex, and age
str(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)

## tibble[,8] [168 × 8] (S3: tbl_df/tbl/data.frame)
##  $ Start Date     : Date[1:168], format: "2020-01-01" "2020-01-01" ...
##  $ End Date       : Date[1:168], format: "2020-01-30" "2020-01-30" ...
##  $ Education      : chr [1:168] "Associate degree or some college"
"Associate degree or some college" "Associate degree or some college"
"Associate degree or some college" ...
##  $ Race           : chr [1:168] "Hispanic" "Hispanic" "Hispanic"
"Hispanic" ...
##  $ Sex            : chr [1:168] "Female" "Female" "Female" "Female" ...
##  $ Age Group      : chr [1:168] "0-17 years" "18-49 years" "50-64 years"
"65 years and over" ...
##  $ COVID-19 Deaths: num [1:168] 0 423 857 1793 0 ...
##  $ Total Deaths   : num [1:168] 2 3117 4153 10225 1 ...

## The following str is for the data frame that is targeting the COVID-19
Deaths by the education and race only
str(COVID_19_Deaths_by_Race_and_Educational)

## tibble[,6] [64 × 6] (S3: tbl_df/tbl/data.frame)
##  $ Start Date     : Date[1:64], format: "2020-01-01" "2020-01-01" ...
##  $ End Date       : Date[1:64], format: "2020-01-30" "2020-01-30" ...
##  $ Education      : chr [1:64] "8th grade or less" "8th grade or less"
"8th grade or less" "8th grade or less" ...
##  $ Race           : chr [1:64] "Hispanic" "Non-Hispanic American Indian or
Alaska Native" "Non-Hispanic Asian" "Non-Hispanic Black" ...
##  $ COVID-19 Deaths: num [1:64] 29157 706 2610 5699 103 ...
##  $ Total Deaths   : num [1:64] 106285 3085 16283 41437 1676 ...

## Last but not least the following str is for the data frame that is
targeting the amount of deaths due to the COVID-19 virus by age but in years
instead of an age group.
str(COVID_19_Deaths_Counts_by_Age_in_Years)

## spec_tbl_df[,7] [172 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Start Date               : Date[1:172], format: "2020-01-01" "2020-01-
01" ...
##  $ End Date                 : Date[1:172], format: "2020-04-30" "2020-04-
30" ...
##  $ Sex                      : chr [1:172] "Male" "Male" "Male" "Male" ...
##  $ Age Years                : chr [1:172] "<1 year" "01 Years" "02 Years"
"03 Years" ...
##  $ Total deaths             : num [1:172] 13525 949 635 515 406 ...
##  $ COVID-19 Deaths          : num [1:172] 40 11 3 3 4 6 1 4 5 6 ...
```

```
##  $ COVID-19 Death Percentage: num [1:172] 0.00296 0.01159 0.00472 0.00583
0.00985 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Data as of` = col_character(),
##   ..   `Start Date` = col_character(),
##   ..   `End Date` = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age Years` = col_character(),
##   ..   `Total deaths` = col_number(),
##   ..   `COVID-19 Deaths` = col_number()
##   .. )
```

After viewing each of these str of data I believe each and every one of these data frames provides knowledgeable data that can help target our key questions.

Research questions and Questions for future steps.

1. Out of the three Pfizer, Moderna, and Johnson and Johnson vaccine for the week of 05/10/2021 which company has the highest number of vaccinations administered?

a. In regard to the number of vaccinations administered which state had the highest rate for each company?

b. In regard to the number of vaccinations administered which state had the lowest rate for each company?

```
summary(COVID_19_Vaccine_Distribution_Pfizer)
```

```
##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:1254        Min.   :2020-01-04   Min.   :   975
##  Class :character   1st Qu.:2020-02-08   1st Qu.: 17550
##  Mode  :character   Median :2020-03-15   Median : 39780
##                     Mean   :2020-04-15   Mean   : 62989
##                     3rd Qu.:2020-04-26   3rd Qu.: 77171
##                     Max.   :2020-12-28   Max.   :644670
##    Company
##  Length:1254
##  Class :character
##  Mode  :character
##
##
##
```

```
summary(COVID_19_Vaccine_Distribution_Moderna)
```

```
##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:1213        Min.   :2020-01-04   Min.   :   700
```

```
##  Class :character   1st Qu.:2020-02-08   1st Qu.: 15600
##  Mode  :character   Median :2020-03-15   Median : 35800
##                     Mean   :2020-04-04   Mean   : 56522
##                     3rd Qu.:2020-04-19   3rd Qu.: 69000
##                     Max.   :2020-12-28   Max.   :672600
##    Company
##  Length:1213
##  Class :character
##  Mode  :character
##
##
##
```

```
summary(COVID_19_Vaccine_Distribution_Johnson)
```

```
##  Jurisdiction       Week of Allocations  1st Dose Allocations
##  Length:486         Min.   :2020-03-01   Min.   :   100
##  Class :character   1st Qu.:2020-03-15   1st Qu.:  3400
##  Mode  :character   Median :2020-03-29   Median :  9750
##                     Mean   :2020-04-03   Mean   : 26018
##                     3rd Qu.:2020-04-12   3rd Qu.: 25300
##                     Max.   :2020-05-10   Max.   :572700
##    Company
##  Length:486
##  Class :character
##  Mode  :character
##
##
##
```

As seen above is the summary of all our data for each companies shot. As shown the the Moderna has the highest doses allocated at 672600 while Pfizer is right behind at 644670 being its max and Johnson last at 572700. I will further research and create new data frames to only look at data for the week of 5/10/2021.

2. what percentage of deaths reported were solely due to Covid-19?

a. When it comes to age and sex which combination had the highest death rate for each sex and age group?

b. As most think the elderly population is the highest at risk for death from the virus but what percentage of deaths in the elderly population was from Covid-19 and what percentage for other illnesses besides Covid-19?

```
summary(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
```

```
##    Start Date            End Date             Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:168
```

```
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##      Race               Sex             Age Group         COVID-19 Deaths
##  Length:168         Length:168         Length:168         Min.   :    0.0
##  Class :character   Class :character   Class :character   1st Qu.:   12.5
##  Mode  :character   Mode  :character   Mode  :character   Median :  133.0
##                                                           Mean   : 2441.9
##                                                           3rd Qu.:  958.5
##                                                           Max.   :76871.0
##   Total Deaths
##  Min.   :     0.0
##  1st Qu.:   159.2
##  Median :  1255.0
##  Mean   : 20479.5
##  3rd Qu.:  8724.8
##  Max.   :670295.0
```

As seen above the max deaths for COVID-19 is 76871.0 while the total deaths is 670295.0. We can easily calculate the percent of COVID-19 deaths by dividing COVID-19 deaths by total deaths. As seen below: Which shows a 11.47% COVID-19 death rate

```
76871.0 /670295.0
```

```
## [1] 0.1146823
```

3. When it comes to race which ethnicity did, we find to have the highest death rate from Covid-19?

a. Did the education level have an effect on the death rate of different ethnicities?

```
summary(COVID_19_Deaths_by_Race_and_Educational)
```

```
##    Start Date            End Date            Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:64
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##      Race           COVID-19 Deaths      Total Deaths
##  Length:64          Min.   :     3.00   Min.   :     21.0
##  Class :character   1st Qu.:    89.25   1st Qu.:    567.2
##  Mode  :character   Median :   684.00   Median :   3907.5
##                     Mean   :  6287.28   Mean   :  53335.8
```

```
##                      3rd Qu.:  4577.50   3rd Qu.:  29872.5
##                      Max.   :117989.00   Max.   :1145594.0
```

4. In the time period of 01/01/2020 to 04/30/2021 did we see an increase in Covid-19 deaths as the age of the individuals rise? If so, did sex have an effect on the number of deaths?

```
summary(COVID_19_Deaths_Counts_by_Age_in_Years)

##    Start Date            End Date              Sex
##  Min.   :2020-01-01   Min.   :2020-04-30   Length:172
##  1st Qu.:2020-01-01   1st Qu.:2020-04-30   Class :character
##  Median :2020-01-01   Median :2020-04-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-04-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-04-30
##  Max.   :2020-01-01   Max.   :2020-04-30
##   Age Years          Total deaths    COVID-19 Deaths
##  Length:172         Min.   :   214   Min.   :    1.00
##  Class :character   1st Qu.:  2296   1st Qu.:   55.25
##  Mode  :character   Median :  8826   Median :  598.00
##                     Mean   : 25897   Mean   : 3289.17
##                     3rd Qu.: 34984   3rd Qu.: 4449.50
##                     Max.   :802843   Max.   :98328.00
##  COVID-19 Death Percentage
##  Min.   :0.002141
##  1st Qu.:0.023516
##  Median :0.085025
##  Mean   :0.078166
##  3rd Qu.:0.126495
##  Max.   :0.158603
```

Questions for future steps.

A question for future steps would be to determine a seemless way to combine all three vaccine data frames together that will make sense with the types of variables each of them have.

What information is not self-evident?

The current information that we have for vaccines rates is not self-evident on which company has the highest rate given per week.

While the information on race is not self-evident on who has the highest death rate from the COVID-19 virus.

Discuss how you plan to uncover new information in the data that is not self-evident.

I am going to pull a certain week from each vaccines data frames and calculate which company had the highest rate of vaccination and which states had the highest from each.

What are different ways you could look at this data to answer the questions you want to answer?

We can look at certain dates to determine the amount of deaths that are due to COVID-19 that week.

Also I can filter by ages to see the mount of deaths that are for each age group.

Seen below is some ways I plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain. That can be viewed with the summary function to shine some light on ome key questions.

```
## The data set titled COVID_19_Deaths_by_Educational_Race_Sex_and_Age is seen below which I will determine how I can slice and dice to create new information.

summary(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)

##    Start Date            End Date           Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:168
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##      Race                Sex             Age Group         COVID-19 Deaths
##  Length:168         Length:168         Length:168         Min.   :    0.0
```

```
##  Class :character   Class :character   Class :character   1st Qu.:   12.5
##  Mode  :character   Mode  :character   Mode  :character   Median :  133.0
##                                                          Mean   : 2441.9
##                                                          3rd Qu.:  958.5
##                                                          Max.   :76871.0
##   Total Deaths
##  Min.   :     0.0
##  1st Qu.:   159.2
##  Median :  1255.0
##  Mean   : 20479.5
##  3rd Qu.:  8724.8
##  Max.   :670295.0
```

```
## I am going to create a new variable to add new insights on the percentage of deaths that are from Covid-19.

COVID_19_Deaths_by_Educational_Race_Sex_and_Age["COVID-19 Death Percentage"] <- COVID_19_Deaths_by_Educational_Race_Sex_and_Age$`COVID-19 Deaths`/ COVID_19_Deaths_by_Educational_Race_Sex_and_Age$`Total Deaths`

## Use summary to search for NAs or other issues with new variable
summary(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
```

```
##    Start Date            End Date            Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:168
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##
##      Race               Sex             Age Group         COVID-19 Deaths
##  Length:168         Length:168         Length:168         Min.   :    0.0
##  Class :character   Class :character   Class :character   1st Qu.:   12.5
##  Mode  :character   Mode  :character   Mode  :character   Median :  133.0
##                                                           Mean   : 2441.9
##                                                           3rd Qu.:  958.5
##                                                           Max.   :76871.0
##
##   Total Deaths      COVID-19 Death Percentage
##  Min.   :     0.0   Min.   :0.00000
##  1st Qu.:   159.2   1st Qu.:0.06112
##  Median :  1255.0   Median :0.12452
##  Mean   : 20479.5   Mean   :0.11994
##  3rd Qu.:  8724.8   3rd Qu.:0.16975
##  Max.   :670295.0   Max.   :0.31181
##                     NA's   :20
```

```
## Next I will get rid of ant NA that have occurred due to the fact there are zeros in both the COVID-19 deaths and total deaths variables
```

```
COVID_19_Deaths_by_Educational_Race_Sex_and_Age[is.na(COVID_19_Deaths_by_Educ
ational_Race_Sex_and_Age)] <- 0.00

## View summary again to show we have no more NAs
summary(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)

##    Start Date            End Date           Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:168
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##      Race               Sex             Age Group         COVID-19 Deaths
##  Length:168         Length:168         Length:168         Min.   :    0.0
##  Class :character   Class :character   Class :character   1st Qu.:   12.5
##  Mode  :character   Mode  :character   Mode  :character   Median :  133.0
##                                                           Mean   : 2441.9
##                                                           3rd Qu.:  958.5
##                                                           Max.   :76871.0
##   Total Deaths      COVID-19 Death Percentage
##  Min.   :     0.0   Min.   :0.00000
##  1st Qu.:   159.2   1st Qu.:0.02515
##  Median :  1255.0   Median :0.11371
##  Mean   : 20479.5   Mean   :0.10566
##  3rd Qu.:  8724.8   3rd Qu.:0.16079
##  Max.   :670295.0   Max.   :0.31181

## Next I will view the COVID_19_Deaths_Counts_by_Age_in_Years data frame to
see if we can create any new variable.
View(COVID_19_Deaths_Counts_by_Age_in_Years)
head(COVID_19_Deaths_Counts_by_Age_in_Years)

## # A tibble: 6 x 7
##   `Start Date` `End Date` Sex   `Age Years` `Total deaths` `COVID-19
Deaths`
##   <date>       <date>     <chr> <chr>               <dbl>
<dbl>
## 1 2020-01-01   2020-04-30 Male  <1 year             13525
40
## 2 2020-01-01   2020-04-30 Male  01 Years              949
11
## 3 2020-01-01   2020-04-30 Male  02 Years              635
3
## 4 2020-01-01   2020-04-30 Male  03 Years              515
3
## 5 2020-01-01   2020-04-30 Male  04 Years              406
4
## 6 2020-01-01   2020-04-30 Male  05 Years              381
```

```
6
## # … with 1 more variable: COVID-19 Death Percentage <dbl>

## As seen in the data frame we can add a new variable that calculates the
deaths that are not related to COVID-19
COVID_19_Deaths_Counts_by_Age_in_Years["Deaths Unrelated to COVID-19"] <-
COVID_19_Deaths_Counts_by_Age_in_Years$`Total deaths`-
COVID_19_Deaths_Counts_by_Age_in_Years$`COVID-19 Deaths`

## As seen in the the data frame now as age goes up the amount of deaths due
to COVID-19 goes up and down as the age goes down with a little spike for 1
year old.
View(COVID_19_Deaths_Counts_by_Age_in_Years)
head(COVID_19_Deaths_Counts_by_Age_in_Years)

## # A tibble: 6 x 8
##   `Start Date` `End Date` Sex   `Age Years` `Total deaths` `COVID-19
Deaths`
##   <date>       <date>     <chr> <chr>               <dbl>
<dbl>
## 1 2020-01-01   2020-04-30 Male  <1 year             13525
40
## 2 2020-01-01   2020-04-30 Male  01 Years              949
11
## 3 2020-01-01   2020-04-30 Male  02 Years              635
3
## 4 2020-01-01   2020-04-30 Male  03 Years              515
3
## 5 2020-01-01   2020-04-30 Male  04 Years              406
4
## 6 2020-01-01   2020-04-30 Male  05 Years              381
6
## # … with 2 more variables: COVID-19 Death Percentage <dbl>,
## #   Deaths Unrelated to COVID-19 <dbl>

## Next I will view the COVID_19_Deaths_by_Race_and_Educational data frame to
see if we can create any new variable.
View(COVID_19_Deaths_by_Race_and_Educational)
head(COVID_19_Deaths_by_Race_and_Educational)

## # A tibble: 6 x 6
##   `Start Date` `End Date` Education  Race        `COVID-19 Death… `Total
Deaths`
##   <date>       <date>     <chr>      <chr>                  <dbl>
<dbl>
## 1 2020-01-01   2020-01-30 8th grade… Hispanic               29157
106285
## 2 2020-01-01   2020-01-30 8th grade… Non-Hispan…              706
3085
## 3 2020-01-01   2020-01-30 8th grade… Non-Hispan…             2610
16283
```

```
## 4 2020-01-01   2020-01-30 8th grade… Non-Hispan…              5699
41437
## 5 2020-01-01   2020-01-30 8th grade… Non-Hispan…               103
1676
## 6 2020-01-01   2020-01-30 8th grade… Non-Hispan…                87
484

## As seen in the data frame we can add new variables that calculates the
deaths that are not related to COVID-19 and percentage of deaths that are due
to COVID-19
## Deaths Unrelated to COVID-19
COVID_19_Deaths_by_Race_and_Educational["Deaths Unrelated to COVID-19"] <-
COVID_19_Deaths_by_Race_and_Educational$`Total Deaths`-
COVID_19_Deaths_by_Race_and_Educational$`COVID-19 Deaths`
## COVID-19 Death Percentage
COVID_19_Deaths_by_Race_and_Educational["COVID-19 Death Percentage"] <-
COVID_19_Deaths_by_Race_and_Educational$`COVID-19
Deaths`/COVID_19_Deaths_by_Race_and_Educational$`Total Deaths`
## Use summary function to view the new added variables
summary(COVID_19_Deaths_by_Race_and_Educational)

##    Start Date            End Date           Education
##  Min.   :2020-01-01   Min.   :2020-01-30   Length:64
##  1st Qu.:2020-01-01   1st Qu.:2020-01-30   Class :character
##  Median :2020-01-01   Median :2020-01-30   Mode  :character
##  Mean   :2020-01-01   Mean   :2020-01-30
##  3rd Qu.:2020-01-01   3rd Qu.:2020-01-30
##  Max.   :2020-01-01   Max.   :2020-01-30
##      Race           COVID-19 Deaths      Total Deaths
##  Length:64          Min.   :     3.00   Min.   :     21.0
##  Class :character   1st Qu.:    89.25   1st Qu.:    567.2
##  Mode  :character   Median :   684.00   Median :   3907.5
##                     Mean   :  6287.28   Mean   :  53335.8
##                     3rd Qu.:  4577.50   3rd Qu.:  29872.5
##                     Max.   :117989.00   Max.   :1145594.0
##  Deaths Unrelated to COVID-19 COVID-19 Death Percentage
##  Min.   :     18.0            Min.   :0.06146
##  1st Qu.:    498.2            1st Qu.:0.11692
##  Median :   3358.5            Median :0.14678
##  Mean   :  47048.5            Mean   :0.14958
##  3rd Qu.:  25657.8            3rd Qu.:0.17917
##  Max.   :1027605.0            Max.   :0.27433

## As we can see we have a lot more deaths that are unrelated to COVID-19
than deaths that are from COVID-19 which will help tie into key questions I
have in regards to COVID-19 and the whole epidemic in general.
```

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

```
## First I will view all data sets for the company Vaccines.
View(COVID_19_Vaccine_Distribution_Pfizer)
head(COVID_19_Vaccine_Distribution_Pfizer)

## # A tibble: 6 x 4
##   Jurisdiction  `Week of Allocations` `1st Dose Allocations` Company
##   <chr>         <date>                                <dbl> <chr>
## 1 Connecticut   2020-05-10                            54990 Pfizer
## 2 Maine         2020-05-10                            21060 Pfizer
## 3 Massachusetts 2020-05-10                           105300 Pfizer
## 4 New Hampshire 2020-05-10                            21060 Pfizer
## 5 Rhode Island  2020-05-10                            16380 Pfizer
## 6 Vermont       2020-05-10                            10530 Pfizer

View(COVID_19_Vaccine_Distribution_Moderna)
head(COVID_19_Vaccine_Distribution_Moderna)

## # A tibble: 6 x 4
##   Jurisdiction  `Week of Allocations` `1st Dose Allocations` Company
##   <chr>         <date>                                <dbl> <chr>
## 1 Connecticut   2020-05-10                            41300 Moderna
## 2 Maine         2020-05-10                            15800 Moderna
## 3 Massachusetts 2020-05-10                            79500 Moderna
## 4 New Hampshire 2020-05-10                            15900 Moderna
## 5 Rhode Island  2020-05-10                            12400 Moderna
## 6 Vermont       2020-05-10                             7500 Moderna

View(COVID_19_Vaccine_Distribution_Johnson)
head(COVID_19_Vaccine_Distribution_Johnson)

## # A tibble: 6 x 4
##   Jurisdiction  `Week of Allocations` `1st Dose Allocations` Company
##   <chr>         <date>                                <dbl> <chr>
## 1 Connecticut   2020-05-10                             6400 Johnson
## 2 Maine         2020-05-10                             2500 Johnson
## 3 Massachusetts 2020-05-10                            12300 Johnson
## 4 New Hampshire 2020-05-10                             2500 Johnson
## 5 Rhode Island  2020-05-10                             2000 Johnson
## 6 Vermont       2020-05-10                             1200 Johnson

## Next I will create new data frames by filtering my data sets to only have
data for 2020-05-10.

## Data frame Date_COVID_19_Vaccine_Distribution_Pfizer
Date_COVID_19_Vaccine_Distribution_Pfizer<-
filter(COVID_19_Vaccine_Distribution_Pfizer,
COVID_19_Vaccine_Distribution_Pfizer$`Week of Allocations`=="2020-05-10")
View(Date_COVID_19_Vaccine_Distribution_Pfizer)
```

```
## created a plot that shows the vaccine rate per state for the week of 2020-05-10 that shows California at the highest vaccinated state for the Pfizer vaccine at 575,640 doses allocated.
ggplot(Date_COVID_19_Vaccine_Distribution_Pfizer, aes(x=`1st Dose Allocations`, y=`Jurisdiction`))+geom_point(aes(fill=`Jurisdiction`))
```


```
## Data frame Date_COVID_19_Vaccine_Distribution_Moderna
Date_COVID_19_Vaccine_Distribution_Moderna<-
filter(COVID_19_Vaccine_Distribution_Moderna,
COVID_19_Vaccine_Distribution_Moderna$`Week of Allocations`=="2020-05-10")
View(Date_COVID_19_Vaccine_Distribution_Moderna)
head(Date_COVID_19_Vaccine_Distribution_Moderna)

## # A tibble: 6 x 4
##   Jurisdiction  `Week of Allocations` `1st Dose Allocations` Company
##   <chr>         <date>                                <dbl> <chr>
## 1 Connecticut   2020-05-10                            41300 Moderna
## 2 Maine         2020-05-10                            15800 Moderna
## 3 Massachusetts 2020-05-10                            79500 Moderna
## 4 New Hampshire 2020-05-10                            15900 Moderna
## 5 Rhode Island  2020-05-10                            12400 Moderna
## 6 Vermont       2020-05-10                             7500 Moderna
```

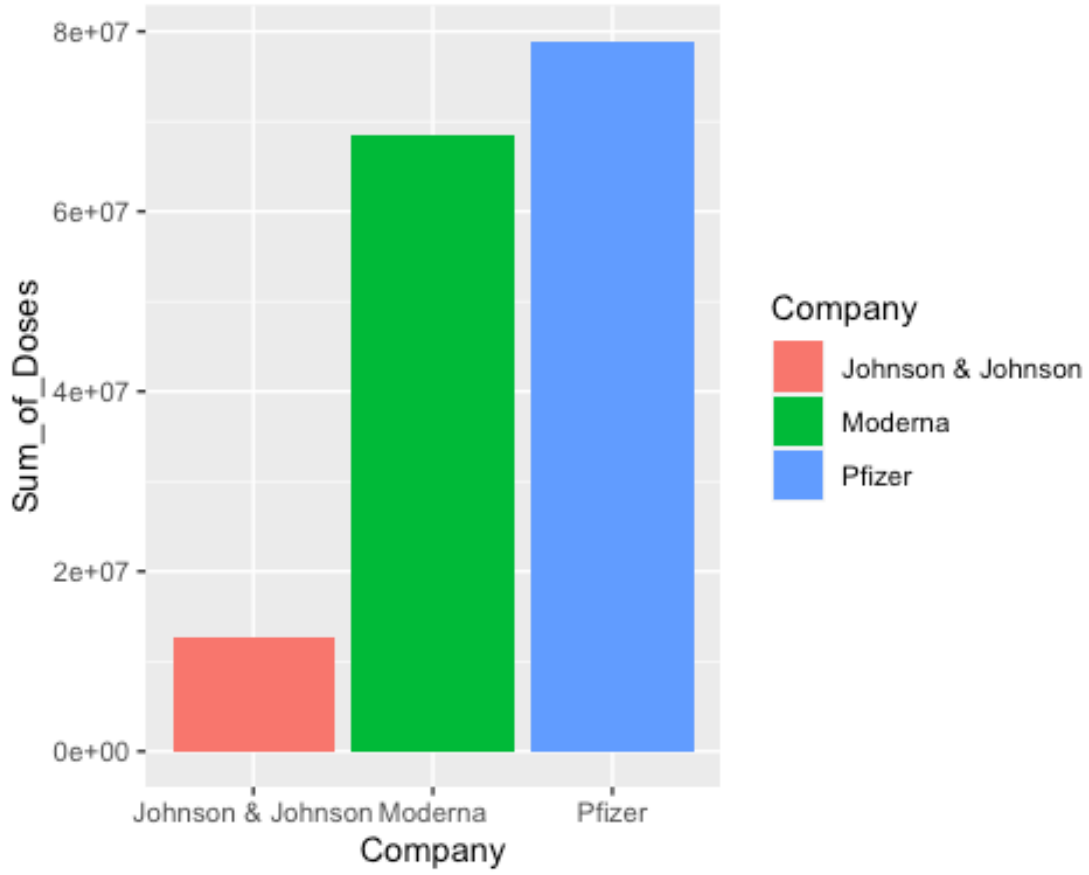
```
## created a plot that shows the vaccine rate per state for the week of 2020-05-10 that shows California at the highest vaccinated state for the Moderna vaccine at 438,100 doses allocated.
ggplot(Date_COVID_19_Vaccine_Distribution_Moderna, aes(x=`1st Dose Allocations`, y=`Jurisdiction`))+geom_point(aes(fill=`Jurisdiction`))
```


```
## Data frame Date_COVID_19_Vaccine_Distribution_Moderna
Date_COVID_19_Vaccine_Distribution_Johnson<-
filter(COVID_19_Vaccine_Distribution_Johnson,
COVID_19_Vaccine_Distribution_Johnson$`Week of Allocations`=="2020-05-10")
View(Date_COVID_19_Vaccine_Distribution_Johnson)
head(Date_COVID_19_Vaccine_Distribution_Johnson)

## # A tibble: 6 x 4
##   Jurisdiction  `Week of Allocations` `1st Dose Allocations` Company
##   <chr>         <date>                                <dbl> <chr>
## 1 Connecticut   2020-05-10                             6400 Johnson
## 2 Maine         2020-05-10                             2500 Johnson
## 3 Massachusetts 2020-05-10                            12300 Johnson
## 4 New Hampshire 2020-05-10                             2500 Johnson
## 5 Rhode Island  2020-05-10                             2000 Johnson
## 6 Vermont       2020-05-10                             1200 Johnson
```

```
## created a plot that shows the vaccine rate per state for the week of 2020-05-10 that shows California at the highest vaccinated state for the Johnson & Johnson vaccine at 67,600 doses allocated.
ggplot(Date_COVID_19_Vaccine_Distribution_Johnson, aes(x=`1st Dose Allocations`, y=`Jurisdiction`))+geom_point(aes(fill=`Jurisdiction`))
```


```
## As seen in these plots we can state that California has reached the highest doses allocated by any state for each companies vaccine while the Pfizer vaccine has rained supreme over the other vaccines.

## Next I am going to create values for the sum of doses for each company for the week of 2020-05-10 and vectors for the amounts and company names.
Pfizer_1st_Dose_Sum <- sum(COVID_19_Vaccine_Distribution_Pfizer$`1st Dose Allocations`)
Moderna_1st_Dose_Sum <-sum(COVID_19_Vaccine_Distribution_Moderna$`1st Dose Allocations`)
Johnson_1st_Dose_Sum <-sum(COVID_19_Vaccine_Distribution_Johnson$`1st Dose Allocations`)
Company <- c("Pfizer", "Moderna", "Johnson & Johnson")
Sum_of_Doses <- c(78987870, 68561180, 12644800)

## Create data frame from the Company and Sum of doses values
Sum_of_Vaccine <- data.frame(Company, Sum_of_Doses)
```

```
View(Sum_of_Vaccine)
head(Sum_of_Vaccine)

##             Company Sum_of_Doses
## 1            Pfizer     78987870
## 2           Moderna     68561180
## 3 Johnson & Johnson     12644800

## Next I will create a bar chart that shows the doses allocated by each company in the United States for the week of 2020-05-10 that shows how many vaccines Moderna and Pfizer administered in comparison to Johnson & Johnson.
ggplot(Sum_of_Vaccine, aes(x=`Company`, y=`Sum_of_Doses`))+geom_col(aes(fill = `Company`))
```



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I plan on using lm() function to create a linear model to predict what the missing data would be for the COVID-19 Deaths, Total Deaths, and COVID-19 Death Percentage variable in the data frame COVID_19_Deaths_by_Educational_Race_Sex_and_Age which will allow me to predict the empty slots by comparing them to similar data I have on these variable.

```
View(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
head(COVID_19_Deaths_by_Educational_Race_Sex_and_Age)
```

```
## # A tibble: 6 x 9
##   `Start Date` `End Date` Education    Race   Sex   `Age Group` `COVID-19
Death…
##   <date>       <date>     <chr>        <chr>  <chr> <chr>
<dbl>
## 1 2020-01-01   2020-01-30 Associate d… Hispa… Fema… 0-17 years
0
## 2 2020-01-01   2020-01-30 Associate d… Hispa… Fema… 18-49 years
423
## 3 2020-01-01   2020-01-30 Associate d… Hispa… Fema… 50-64 years
857
## 4 2020-01-01   2020-01-30 Associate d… Hispa… Fema… 65 years a…
1793
## 5 2020-01-01   2020-01-30 Associate d… Hispa… Male  0-17 years
0
## 6 2020-01-01   2020-01-30 Associate d… Hispa… Male  18-49 years
737
## # … with 2 more variables: Total Deaths <dbl>, COVID-19 Death Percentage
<dbl>
```

Questions for future steps.

What do you not know how to do right now that you need to learn to answer your questions?

I am going to need to brush up on how to extract uneeded characters in my values in my COVID-19 Death rates data frames especially in the age and race variables to clean them up.

How can you incorporate the pipe (%>%) operator to make your code more efficient?

After working on this assignment I determined I can incorporate pipe in my project to help me make my filtered data more ledgable and easier to follow.