Covid 19 Vaccine Analysis Milestone 2

October 2, 2024

0.0.1 Cleaning Flat File

[2]:

```
import pandas as pd
import numpy as np
import requests
from bs4 import BeautifulSoup
```

[3]:

```
# read file
```

[4]:

```
covid_df = pd.read_csv("/Users/yuhang/Desktop/DSC540/term_project/
 ↪COVID-19_Vaccinations_in_the_US.csv")
covid_df.head(10)
```

[4]:

```
         Date  MMWR_week Location  Distributed  Distributed_Janssen  \
0  05/10/2023         19       NE      5481710               152400
1  05/10/2023         19       LA     10282120               330500
2  05/10/2023         19       GA     28727475               869100
3  05/10/2023         19       WY      1281755                49300
4  05/10/2023         19       CO     17769135               501900
5  05/10/2023         19       PA     42895735              1569200
6  05/10/2023         19       DE      3352025               102600
7  05/10/2023         19       OR     14222125               466700
8  05/10/2023         19       WI     16474175               457200
9  05/10/2023         19       ND      1910860                53600

   Distributed_Moderna  Distributed_Pfizer  Distributed_Novavax  \
0              1647380             2905630               7400.0
1              3807980             5164550              10100.0
2              9763000            14773655              43400.0
3               490040              585605               3700.0
4              5402640             9029715              43600.0
5             13941120            21178525              87400.0
6              1071000             1651775               5800.0
7              4482360             7060535              25900.0
8              5144600             8183105              22900.0
9               600920              990720               2800.0

   Distributed_Unk_Manuf  Dist_Per_100K  …  Dist_Bivalent_PFR  \
0                      0          283379 …            575480.0
1                      0          221178 …            640590.0
2                      0          270569 …           2255000.0
3                      0          221466 …            102510.0
4                      0          308560 …           2033760.0
5                      0          335071 …           4206470.0
6                      0          344234 …            357750.0
7                      0          337198 …           1631890.0
8                      0          282943 …           1815090.0
9                      0          250749 …            203500.0

   Dist_Bivalent_MOD  Bivalent_Booster_5Plus  Bivalent_Booster_5Plus_Pop_Pct  \
0           193420.0                340508.0                            18.9   
1           328400.0                359506.0                             8.3   
2          1023320.0               1126791.0                            11.3   
3            50600.0                 65920.0                            12.1   
4           757520.0               1268501.0                            23.4   
5          1913020.0               2404885.0                            19.9   
6           163100.0                206298.0                            22.4   
7           554740.0                944587.0                            23.7   
8           851280.0               1330586.0                            24.2   
9            59320.0                120653.0                            17.0   

   Bivalent_Booster_12Plus  Bivalent_Booster_12Plus_Pop_Pct  \
0                 332054.0                             20.5   
1                 356199.0                              9.1   
2                1103841.0                             12.3   
3                  64919.0                             13.2   
4                1229981.0                             24.9   
5                2349980.0                             21.2   
6                 202637.0                             24.1   
7                 913832.0                             25.1   
8                1298038.0                             26.0   
9                 117294.0                             18.4   

   Bivalent_Booster_18Plus  Bivalent_Booster_18Plus_Pop_Pct  \
0                 319161.0                             21.9   
1                 349934.0                              9.8   
2                1067468.0                             13.2   
3                  63244.0                             14.2   
4                1176686.0                             26.2   
5                2271654.0                             22.3   
6                 196313.0                             25.5   
7                 874306.0                             26.1   
8                1251270.0                             27.5   
9                 112872.0                             19.4   
```

```
   Bivalent_Booster_65Plus  Bivalent_Booster_65Plus_Pop_Pct
0                 151146.0                             48.4
1                 184844.0                             24.9
2                 476297.0                             31.4
3                  34832.0                             35.1
4                 462990.0                             55.0
5                1105455.0                             46.2
6                 101880.0                             53.9
7                 371182.0                             48.5
8                 600688.0                             59.1
9                  56881.0                             47.5

[10 rows x 109 columns]
```

[5]:
```
# covid_df.describe()
```

[6]:
```
# covid_df.columns.values
```

[7]:
```
# step 1 check if there is null values
```

[8]:
```
covid_df.isnull().sum().values
```

[8]:
```
array([    0,     0,     0,     0,     0,     0,     0, 35800,     0,
           0,   448,     0,     0,     0,     0,   448,     0,     0,
           0,     0,     0,     0, 35807,     3,     0,   448,     0,
           0,     0,     0,     0,     0,   448,   448,     0,     0,
           0,     0,     0,     0,     0,     0,   448,   448,     0,
           0,     0,     0,     0,     0,     0,     0,     0, 35808,
           4, 21016, 21016, 21016, 21020,     0,     0,     0,     4,
           0,     0,     0,     4,     0,     0,     0,     9, 16348,
         325, 35544, 35544, 26456, 26456,   325,   325,   325,     0,
         325,   325,   325,   325,   327,   331, 38385, 31896, 31896,
       31896, 31896, 31905, 31896, 31896, 31907, 36248, 36312, 36312,
       36312, 36312, 36568, 36568, 36504, 36504, 36504, 36504, 36504,
       36504])
```

[9]:
```
covid_df.shape
```

[9]:
```
(38488, 109)
```

[10]:
```
# step 2 select only some of them out of 109 columns
```

[11]:
```
covid_new = covid_df[["Date", "Location", "Distributed", "Administered",␣
 ↪"Administered_5Plus", "Administered_12Plus",\
           "Administered_18Plus", "Administered_65Plus",␣
 ↪"Administered_Moderna",\
           "Additional_Doses", "Additional_Doses_5Plus",␣
 ↪"Additional_Doses_12Plus",\
           "Additional_Doses_18Plus", "Additional_Doses_65Plus",␣
 ↪"Second_Booster"]]
 covid_new.head(10)
```

```
[11]:         Date Location  Distributed  Administered  Administered_5Plus  \
      0  05/10/2023       NE      5481710       3822190           3793971.0
      1  05/10/2023       LA     10282120       6961453           6945414.0
      2  05/10/2023       GA     28727475      17124791          17045184.0
      3  05/10/2023       WY      1281755        854132            851464.0
      4  05/10/2023       CO     17769135      13033446          12899729.0
      5  05/10/2023       PA     42895735      27586432          27360998.0
      6  05/10/2023       DE      3352025       2169125           2157007.0
      7  05/10/2023       OR     14222125       9399175           9326386.0
      8  05/10/2023       WI     16474175      12444016          12347025.0
      9  05/10/2023       ND      1910860       1314469           1302373.0

         Administered_12Plus  Administered_18Plus  Administered_65Plus  \
      0              3647301              3412154              1117112
      1              6796682              6443990              2090638
      2             16545894             15542310              4409764
      3               831727               790769               284622
      4             12396074             11614071              3172390
      5             26407167             24920539              8490730
      6              2088591              1964981               709063
      7              8983058              8453948              2698104
      8             11929874             11264241              3948468
      9              1253917              1185548               387008

         Administered_Moderna  Additional_Doses  Additional_Doses_5Plus  \
      0               1240872          718168.0                716720.0
      1               2685630         1110217.0               1110101.0
      2               6170654         2705136.0               2703444.0
      3                347522          145600.0                145579.0
      4               4490965         2460212.0               2456349.0
      5               9697012         4493396.0               4484903.0
      6                747546          369899.0                369753.0
      7               3129879         1812982.0               1811925.0
      8               4216126         2432044.0               2429044.0
      9                430115          219540.0                219501.0

         Additional_Doses_12Plus  Additional_Doses_18Plus  Additional_Doses_65Plus  \
      0                 699462.0                 664415.0                 240367.0
      1                1103756.0                1076600.0                 450792.0
      2                2666367.0                2567417.0                 892658.0
      3                 143988.0                 139599.0                  62044.0
```

```
4                2391387.0                2265625.0                 662890.0
5                4394827.0                4213266.0                1683048.0
6                 363440.0                 347221.0                 144617.0
7                1765406.0                1676283.0                 582631.0
8                2374846.0                2270627.0                 847457.0
9                 214661.0                 206044.0                  81722.0

   Second_Booster
0             NaN
1             NaN
2             NaN
3             NaN
4             NaN
5             NaN
6             NaN
7             NaN
8             NaN
9             NaN
```

[12]:

```
# step 3 drop unwanted rows for analysis on first doses
```

[13]:

```
df_first = covid_new.dropna(subset=['Administered'])
df_first.shape
```

[13]:

```
(38488, 15)
```

[14]:

```
df_first.tail()
```

[14]:

```
             Date Location  Distributed  Administered  Administered_5Plus  \
38483  12/13/2020        AS         3900             0                 0.0
38484  12/13/2020        VI          975             0                 0.0
38485  12/13/2020        MP         4875             0                 0.0
38486  12/13/2020        US        13650             0                 0.0
38487  12/13/2020        GU         3900             0                 0.0

       Administered_12Plus  Administered_18Plus  Administered_65Plus  \
38483                    0                    0                    0
38484                    0                    0                    0
38485                    0                    0                    0
38486                    0                    0                    0
38487                    0                    0                    0

       Administered_Moderna  Additional_Doses  Additional_Doses_5Plus  \
38483                     0               NaN                     NaN
38484                     0               NaN                     NaN
38485                     0               NaN                     NaN
38486                     0               NaN                     NaN
```

```
38487                     0                NaN                       NaN

       Additional_Doses_12Plus  Additional_Doses_18Plus  \
38483                      NaN                      0.0
38484                      NaN                      0.0
38485                      NaN                      0.0
38486                      NaN                      0.0
38487                      NaN                      0.0

       Additional_Doses_65Plus  Second_Booster
38483                      0.0             NaN
38484                      0.0             NaN
38485                      0.0             NaN
38486                      0.0             NaN
38487                      0.0             NaN
```

[15]:

```
df_first.columns.values
```

[15]:

```
array(['Date', 'Location', 'Distributed', 'Administered',
       'Administered_5Plus', 'Administered_12Plus', 'Administered_18Plus',
       'Administered_65Plus', 'Administered_Moderna', 'Additional_Doses',
       'Additional_Doses_5Plus', 'Additional_Doses_12Plus',
       'Additional_Doses_18Plus', 'Additional_Doses_65Plus',
       'Second_Booster'], dtype=object)
```

[16]:

```
df_first.isnull().sum().values
```

[16]:

```
array([    0,     0,     0,     0,   448,     0,     0,     0,     0,
       16348, 35544, 26456,   325,   325, 38385])
```

[17]:

```
# step 4 remove fuzzy read strings in "location" column
```

[18]:

```
df_first['Location'].unique()
```

[18]:

```
array(['NE', 'LA', 'GA', 'WY', 'CO', 'PA', 'DE', 'OR', 'WI', 'ND', 'TX',
       'MN', 'UT', 'SC', 'DC', 'NC', 'WA', 'SD', 'PR', 'RI', 'IA', 'FM',
       'PW', 'NV', 'KY', 'VI', 'WV', 'VA2', 'ME', 'ID', 'BP2', 'MP', 'US',
       'IH2', 'MS', 'IL', 'KS', 'MH', 'FL', 'MO', 'GU', 'VT', 'CT', 'OH',
       'NJ', 'DD2', 'TN', 'CA', 'MT', 'IN', 'NY', 'AL', 'VA', 'MD', 'AR',
       'HI', 'OK', 'NH', 'AZ', 'MI', 'AS', 'AK', 'MA', 'NM', 'RP', 'LTC'],
      dtype=object)
```

[19]:

```
states = ["AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DE", "FL", "GA", "HI",␣
 ↪"IA",
    "ID", "IL", "IN", "KS", "KY", "LA", "MA", "MD", "ME", "MI", "MN", "MO",
    "MS", "MT", "NC", "ND", "NE", "NH", "NJ", "NM", "NV", "NY", "OH", "OK",
    "OR", "PA", "RI", "SC", "SD", "TN", "TX", "UT", "VA", "VT", "WA", "WI",
    "WV", "WY", "DC", "AS", "GU", "MP", "PR", "VI"]
```

[20]:
```
len(states)
```

[20]:
```
56
```

[21]:
```
df_first1 = df_first[df_first['Location'].isin(states)]
```

[22]:
```
df_first1.shape
```

[22]:
```
(33436, 15)
```

[23]:
```
# df_first1['Location'].unique()
```

[24]:
```
# step 5 change header since the header about age is a bit confusing
```

[25]:
```
df_first1 = df_first1.rename(columns={'Administered_5Plus':␣
  ↪'Administered_5-12',\
                                       'Administered_12Plus':␣
  ↪'Administered_12-18',\
                                       'Administered_18Plus':␣
  ↪'Administered_18-65',
                                       'Additional_Doses_5Plus':␣
  ↪'Additional_Doses_5-12',\
                                       'Additional_Doses_12Plus':␣
  ↪'Additional_Doses_12-18',\
                                       'Additional_Doses_18Plus':␣
  ↪'Additional_Doses_18-65'})
```

[26]:
```
df_first1.head(10)
```

[26]:
```
         Date Location  Distributed  Administered  Administered_5-12  \
0  05/10/2023       NE      5481710       3822190          3793971.0
1  05/10/2023       LA     10282120       6961453          6945414.0
2  05/10/2023       GA     28727475      17124791         17045184.0
3  05/10/2023       WY      1281755        854132           851464.0
4  05/10/2023       CO     17769135      13033446         12899729.0
5  05/10/2023       PA     42895735      27586432         27360998.0
6  05/10/2023       DE      3352025       2169125          2157007.0
7  05/10/2023       OR     14222125       9399175          9326386.0
8  05/10/2023       WI     16474175      12444016         12347025.0
9  05/10/2023       ND      1910860       1314469          1302373.0

   Administered_12-18  Administered_18-65  Administered_65Plus  \
0             3647301             3412154              1117112
1             6796682             6443990              2090638
2            16545894            15542310              4409764
3              831727              790769               284622
4            12396074            11614071              3172390
```

```
5            26407167            24920539              8490730
6             2088591             1964981               709063
7             8983058             8453948              2698104
8            11929874            11264241              3948468
9             1253917             1185548               387008

   Administered_Moderna  Additional_Doses  Additional_Doses_5-12  \
0               1240872          718168.0               716720.0
1               2685630         1110217.0              1110101.0
2               6170654         2705136.0              2703444.0
3                347522          145600.0               145579.0
4               4490965         2460212.0              2456349.0
5               9697012         4493396.0              4484903.0
6                747546          369899.0               369753.0
7               3129879         1812982.0              1811925.0
8               4216126         2432044.0              2429044.0
9                430115          219540.0               219501.0

   Additional_Doses_12-18  Additional_Doses_18-65  Additional_Doses_65Plus  \
0                699462.0                664415.0                 240367.0
1               1103756.0               1076600.0                 450792.0
2               2666367.0               2567417.0                 892658.0
3                143988.0                139599.0                  62044.0
4               2391387.0               2265625.0                 662890.0
5               4394827.0               4213266.0                1683048.0
6                363440.0                347221.0                 144617.0
7               1765406.0               1676283.0                 582631.0
8               2374846.0               2270627.0                 847457.0
9                214661.0                206044.0                  81722.0

   Second_Booster
0             NaN
1             NaN
2             NaN
3             NaN
4             NaN
5             NaN
6             NaN
7             NaN
8             NaN
9             NaN
```

[27]:
```
# step 6 merge columns
```

[28]:
```
# insert column for first doses for age 0-18
```

[29]:

```
df_first1.insert(4, "Administered_0-18", df_first1['Administered_5-12'] +␣
 ↪df_first1['Administered_12-18'])
```

[30]:

```
df_first1.head(10)
```

[30]:

```
         Date Location  Distributed  Administered  Administered_0-18  \
0  05/10/2023       NE      5481710       3822190          7441272.0
1  05/10/2023       LA     10282120       6961453         13742096.0
2  05/10/2023       GA     28727475      17124791         33591078.0
3  05/10/2023       WY      1281755        854132          1683191.0
4  05/10/2023       CO     17769135      13033446         25295803.0
5  05/10/2023       PA     42895735      27586432         53768165.0
6  05/10/2023       DE      3352025       2169125          4245598.0
7  05/10/2023       OR     14222125       9399175         18309444.0
8  05/10/2023       WI     16474175      12444016         24276899.0
9  05/10/2023       ND      1910860       1314469          2556290.0

   Administered_5-12  Administered_12-18  Administered_18-65  \
0          3793971.0             3647301             3412154
1          6945414.0             6796682             6443990
2         17045184.0            16545894            15542310
3           851464.0              831727              790769
4         12899729.0            12396074            11614071
5         27360998.0            26407167            24920539
6          2157007.0             2088591             1964981
7          9326386.0             8983058             8453948
8         12347025.0            11929874            11264241
9          1302373.0             1253917             1185548

   Administered_65Plus  Administered_Moderna  Additional_Doses  \
0              1117112               1240872          718168.0
1              2090638               2685630         1110217.0
2              4409764               6170654         2705136.0
3               284622                347522          145600.0
4              3172390               4490965         2460212.0
5              8490730               9697012         4493396.0
6               709063                747546          369899.0
7              2698104               3129879         1812982.0
8              3948468               4216126         2432044.0
9               387008                430115          219540.0

   Additional_Doses_5-12  Additional_Doses_12-18  Additional_Doses_18-65  \
0               716720.0                699462.0                664415.0
1              1110101.0               1103756.0               1076600.0
2              2703444.0               2666367.0               2567417.0
3               145579.0                143988.0                139599.0
4              2456349.0               2391387.0               2265625.0
```

```
5         4484903.0                4394827.0                4213266.0
6          369753.0                 363440.0                 347221.0
7         1811925.0                1765406.0                1676283.0
8         2429044.0                2374846.0                2270627.0
9          219501.0                 214661.0                 206044.0

   Additional_Doses_65Plus  Second_Booster
0                 240367.0             NaN
1                 450792.0             NaN
2                 892658.0             NaN
3                  62044.0             NaN
4                 662890.0             NaN
5                1683048.0             NaN
6                 144617.0             NaN
7                 582631.0             NaN
8                 847457.0             NaN
9                  81722.0             NaN
```

```
[31]: # insert column for addition doses for age 0-18
```

```
[32]: df_first1.insert(11, "Additional_Doses_0-18",␣
       ↪df_first1['Additional_Doses_5-12'] + df_first1['Additional_Doses_12-18'])
```

```
[33]: # drop unnecessary columns
```

```
[34]: df_first1 = df_first1.drop(columns=['Administered_5-12', 'Administered_12-18',␣
       ↪'Additional_Doses_5-12', 'Additional_Doses_12-18'])
```

```
[35]: df_first1
```

```
[35]:              Date Location  Distributed  Administered  Administered_0-18  \
      0      05/10/2023       NE      5481710       3822190          7441272.0
      1      05/10/2023       LA     10282120       6961453         13742096.0
      2      05/10/2023       GA     28727475      17124791         33591078.0
      3      05/10/2023       WY      1281755        854132          1683191.0
      4      05/10/2023       CO     17769135      13033446         25295803.0
      …             …      …          …           …                 …
      38481  12/14/2020       MA         5850             0                0.0
      38483  12/13/2020       AS         3900             0                0.0
      38484  12/13/2020       VI          975             0                0.0
      38485  12/13/2020       MP         4875             0                0.0
      38487  12/13/2020       GU         3900             0                0.0

             Administered_18-65  Administered_65Plus  Administered_Moderna  \
      0                 3412154              1117112               1240872
      1                 6443990              2090638               2685630
      2                15542310              4409764               6170654
```

```
3                  790769               284622                347522
4                11614071              3172390               4490965
…                   …                  …                   …
38481                   0                    0                     0
38483                   0                    0                     0
38484                   0                    0                     0
38485                   0                    0                     0
38487                   0                    0                     0

       Additional_Doses  Additional_Doses_0-18  Additional_Doses_18-65  \
0              718168.0              1416182.0                664415.0
1             1110217.0              2213857.0               1076600.0
2             2705136.0              5369811.0               2567417.0
3              145600.0               289567.0                139599.0
4             2460212.0              4847736.0               2265625.0
…                  …                    …                     …
38481               NaN                    NaN                     0.0
38483               NaN                    NaN                     0.0
38484               NaN                    NaN                     0.0
38485               NaN                    NaN                     0.0
38487               NaN                    NaN                     0.0

       Additional_Doses_65Plus  Second_Booster
0                     240367.0             NaN
1                     450792.0             NaN
2                     892658.0             NaN
3                      62044.0             NaN
4                     662890.0             NaN
…                        …             …
38481                      0.0             NaN
38483                      0.0             NaN
38484                      0.0             NaN
38485                      0.0             NaN
38487                      0.0             NaN

[33436 rows x 13 columns]
```

[36]:

```
df_first1.tail()
```

[36]:

```
             Date Location  Distributed  Administered  Administered_0-18  \
38481  12/14/2020       MA         5850             0                0.0
38483  12/13/2020       AS         3900             0                0.0
38484  12/13/2020       VI          975             0                0.0
38485  12/13/2020       MP         4875             0                0.0
38487  12/13/2020       GU         3900             0                0.0

       Administered_18-65  Administered_65Plus  Administered_Moderna  \
38481                   0                    0                     0
38483                   0                    0                     0
38484                   0                    0                     0
38485                   0                    0                     0
38487                   0                    0                     0

       Additional_Doses  Additional_Doses_0-18  Additional_Doses_18-65  \
38481               NaN                    NaN                     0.0
38483               NaN                    NaN                     0.0
38484               NaN                    NaN                     0.0
38485               NaN                    NaN                     0.0
38487               NaN                    NaN                     0.0

       Additional_Doses_65Plus  Second_Booster
38481                      0.0             NaN
38483                      0.0             NaN
38484                      0.0             NaN
38485                      0.0             NaN
38487                      0.0             NaN
```

```
[37]: # drop unnecessary rows
```

```
[38]: df_second = df_first1.dropna(subset=['Additional_Doses'])
      df_second.shape
```

```
[38]: (19320, 13)
```

Several edits and reorganization of the data has been made to the raw data file. First of all, T checked if there is null values in the dataframe, and select the columns I want to use with my analysis from 109 columns; a few data column names has been revised to avoid ambiguity; new column is created to sum up two groups into one to align with other data sources; Lastly I dropped some datasets that has no values in 'Additional_Dose' column to form a dataframe maily focus on second dose for future analysis.

This COVID-19 vaccination data is directly shared by CDC to the public domain, so it's from a highly creditable source. Apparently further handling of the data will follow guidelines and recommendations of COVID-19 vaccination, including data reporting requirements for healthcare providers and public health agencies. Also, we must adhere to data security standards to protect against unauthorized access, use, or disclosure of sensitive information.

The data transformation targets for creating more clear demonstration of the information, with most relevant aspects maintained. The transformation enables a better alignment between this data frame with other data sources. The only one assumption being made in this transformation is about the age groups, we expect the number of infants(kids under 5 years old) who takes COVID-19 vaccination is for the most part small enough compared with other ages groups, so that when we created the new "additional_Doses_0-18" column, we added up the 5Plus (5-12 years old) and 12 Plus (12-18 years old) column and expected it will be substaintially representing the the 0-18 age group.

However we still missing the information about vaccination for age group under 5 in our 0-18 age group. That might lead to the inaccuracy of the analysis. It wouldn't be a big problem, as we all have known that the group that had the vaccination under age 5 are slight small, that might make a big difference in the data analysis process. No assumption was made during my data cleaning process, I would like to see some amazing results in the next analysis steps.

[39]:
```
file_path = "df_first1.csv"
df_first1.to_csv(file_path, index=False)
```

[ ]:

[ ]:

[ ]:

[ ]: