CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

Module Title:	Data Exploration & Preparation
Assessment Title:	CA1 Project
Lecturer Name:	Dr. Muhammad Iqbal
Student Full Name:	Mateus Fonseca Campos
Student Number:	2023327
Assessment Due Date:	03/12/2023
Date of Submission:	03/12/2023

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

COVID-19 Vaccination Trends in the United States

Continuous Assessment 1

Word Count: 2267

Mateus Fonseca Campos

2023327

Contents

Figures

Tables

Introduction

This assignment uses a COVID-19 dataset to perform Data Exploration and Preparation tasks with the intention of identifying trends related to the topic.

The paper is divided into three major sections:

1. **Data Preparation:** data cleaning and preprocessing.
2. **Exploratory Data Analysis:** data exploration and understanding.
3. **Principal Component Analysis:** dimensionality reduction and further exploration.

Data Preparation

The dataset analysed in this assignment is the COVID-19 Vaccination Trends in the United States, National and Jurisdictional (Centers for Disease Control and Prevention, 2023). It has 88,560 rows and 29 columns.

The table below is the definition of each column in the dataset as per the publisher's website:

Column Name	Description	Type
Date	Date data are reported on CDC COVID Data Tracker	Date & Time
date_type	Date of administration or date reported by CDC on COVID Tracker	Plain Text
MMWR_week	The week of the epidemiologic year as defined by the Morbidity and Mortality Weekly Report (https://ndc.services.cdc.gov/wp-content/uploads/MMWR_week_overview.pdf).	Number
Location	State/Territory/Federal Entity	Plain Text
Administered_Daily	Total number of administered doses by date of administration.	Number
Administered_Cumulative	Cumulative number of reported doses administered by date of administration	Number
Administered_7_Day_Rolling_Average	7-day moving average of the daily doses administered by date of administration	Number
Admin_Dose_1_Daily	Total number of dose 1 administations by date of administration	Number
Admin_Dose_1_Cumulative	Cumulative number of people with at least one dose of any vaccine by date of administration.	Number
Admin_Dose_1_Day_Rolling_Average	7-day moving average count of people with at least one dose of any vaccine by date of administration	Number
Administered_Dose1_Pop_Pct	Percent of population with at least one dose based on the jurisdiction where recipient lives	Number
Administered_daily_change_report	Change between the cumulative number of doses administered on a given day and the previous day by date of report	Number
Administered_daily_change_report_7dayroll	7-day moving average of the daily change based by date of report	Number
Series_Complete_Daily	Daily total count of people with a completed	Number

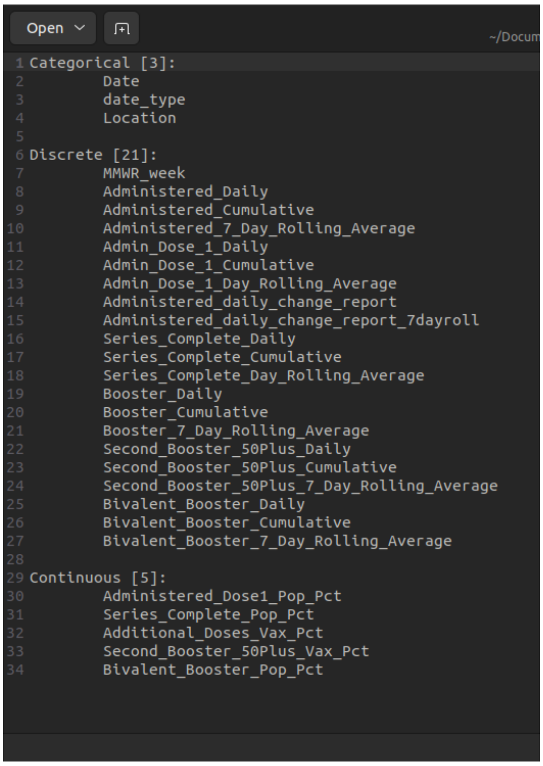
	primary series by date of administration	
Series_Complete_Cumulative	Cumulative total of people with a completed primary series by date of administration	Number
Series_Complete_Day_Rolling_Average	7-day moving average count of people with a completed primary series by date of administration	Number
Series_Complete_Pop_Pct	Percent of people with a completed primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction where recipient lives	Number
Booster_Daily	Daily total count of people who have completed a primary series and have received a booster (or additional) dose by date of administration	Number
Booster_Cumulative	Cumulative total of people who have completed a primary series and have received a booster (or additional) dose by date of administration	Number
Booster_7_Day_Rolling_Average	7-day moving average count of people who have completed a primary series and have received a booster (or additional) dose by date of administration	Number
Additional_Doses_Vax_Pct	Percent of people who have completed a primary series and have received a booster (or additional) dose.	Number
Second_Booster_50Plus_Daily	Daily count of people ages 50+ receiving a second booster dose	Number
Second_Booster_50Plus_Cumulative	Cumulative total of people ages 50+ who have received a second booster dose	Number
Second_Booster_50Plus_7_Day_Rolling_Average	7-day moving average count of people ages 50+ who have received a second booster dose	Number
Second_Booster_50Plus_Vax_Pct	Percent of people ages 50+ with a first booster dose who received a second booster dose	Number
Bivalent_Booster_Daily	Total number of administered bivalent booster doses by date of administration	Plain Text
Bivalent_Booster_Cumulative	Cumulative number of reported bivalent booster doses administered by date of administration	Plain Text
Bivalent_Booster_7_Day_Rolling_Average	7-day moving average of the daily bivalent booster doses administered by date of administration	Plain Text
Bivalent_Booster_Pop_Pct	Percent of population with a bivalent booster	Plain Text

	dose based on the jurisdiction where recipient lives

Table 1: Definition of the columns in the dataset (ibid.)

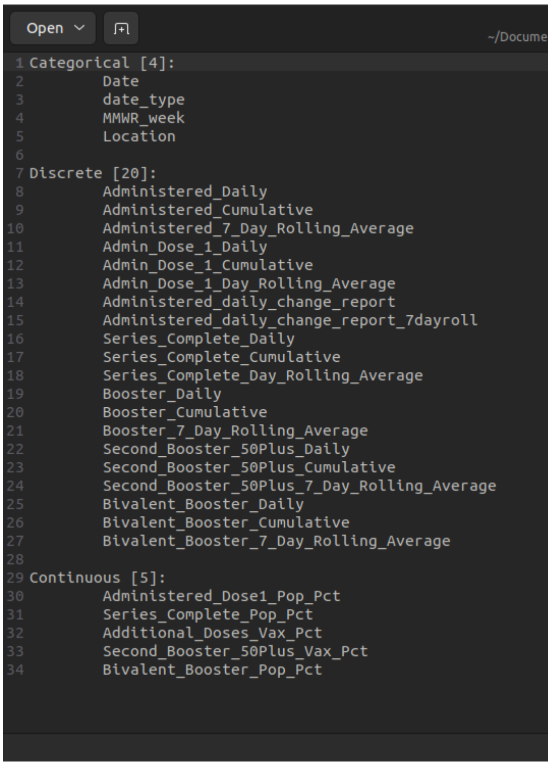
Variable classification

The images below show how the variables in the dataset can be classified into categorical, discrete or continuous. The figure on the left has the column MMWR_week as a discrete numeric variable. This variable was converted to factor so that it could be treated as a variable intended for categorization rather than calculations:



```
Categorical [3]:
        Date
        date_type
        Location

Discrete [21]:
        MMWR_week
        Administered_Daily
        Administered_Cumulative
        Administered_7_Day_Rolling_Average
        Admin_Dose_1_Daily
        Admin_Dose_1_Cumulative
        Admin_Dose_1_Day_Rolling_Average
        Administered_daily_change_report
        Administered_daily_change_report_7dayroll
        Series_Complete_Daily
        Series_Complete_Cumulative
        Series_Complete_Day_Rolling_Average
        Booster_Daily
        Booster_Cumulative
        Booster_7_Day_Rolling_Average
        Second_Booster_50Plus_Daily
        Second_Booster_50Plus_Cumulative
        Second_Booster_50Plus_7_Day_Rolling_Average
        Bivalent_Booster_Daily
        Bivalent_Booster_Cumulative
        Bivalent_Booster_7_Day_Rolling_Average

Continuous [5]:
        Administered_Dose1_Pop_Pct
        Series_Complete_Pop_Pct
        Additional_Doses_Vax_Pct
        Second_Booster_50Plus_Vax_Pct
        Bivalent_Booster_Pop_Pct
```

Figure 1: Classification of the variables in the dataset (MMWR_week as number)



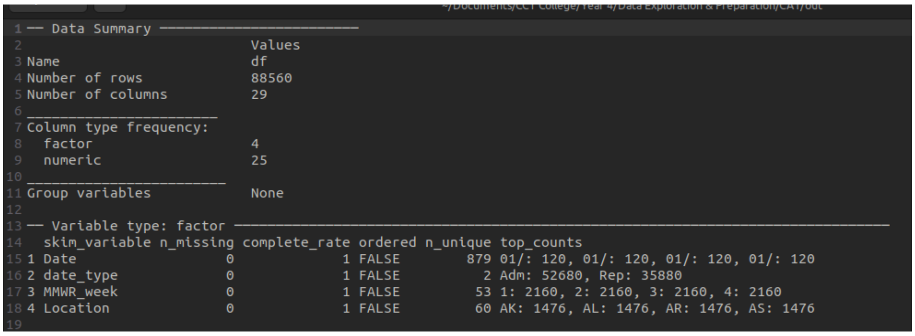
```
Categorical [4]:
        Date
        date_type
        MMWR_week
        Location

Discrete [20]:
        Administered_Daily
        Administered_Cumulative
        Administered_7_Day_Rolling_Average
        Admin_Dose_1_Daily
        Admin_Dose_1_Cumulative
        Admin_Dose_1_Day_Rolling_Average
        Administered_daily_change_report
        Administered_daily_change_report_7dayroll
        Series_Complete_Daily
        Series_Complete_Cumulative
        Series_Complete_Day_Rolling_Average
        Booster_Daily
        Booster_Cumulative
        Booster_7_Day_Rolling_Average
        Second_Booster_50Plus_Daily
        Second_Booster_50Plus_Cumulative
        Second_Booster_50Plus_7_Day_Rolling_Average
        Bivalent_Booster_Daily
        Bivalent_Booster_Cumulative
        Bivalent_Booster_7_Day_Rolling_Average

Continuous [5]:
        Administered_Dose1_Pop_Pct
        Series_Complete_Pop_Pct
        Additional_Doses_Vax_Pct
        Second_Booster_50Plus_Vax_Pct
        Bivalent_Booster_Pop_Pct
```

Figure 2: Classification of the variables in the dataset (MMWR_week as factor)

Statistical parameters

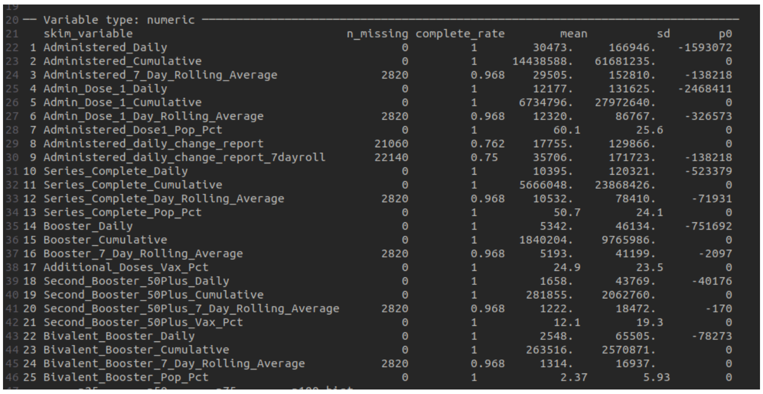
The images below show a more detailed description of each variable in the dataset. It can be seen that some of the numerical variables have missing values, as well as negative minimum values. Given the nature of the dataset, negative values should not be present, since all the numeric values represent either proportion or count.

Figures 3, 4 and 5 show the details before treatment:



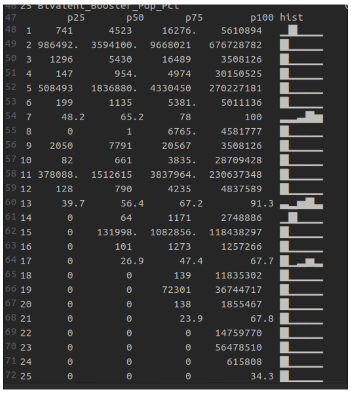
```
~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/out
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             88560
Number of columns          29
_______________________
Column type frequency:
  factor                   4
  numeric                  25
________________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 Date                  0             1 FALSE        879 01/: 120, 01/: 120, 01/: 120, 01/: 120
2 date_type             0             1 FALSE          2 Adm: 52680, Rep: 35880
3 MMWR_week             0             1 FALSE         53 1: 2160, 2: 2160, 3: 2160, 4: 2160
4 Location              0             1 FALSE         60 AK: 1476, AL: 1476, AR: 1476, AS: 1476
```

Figure 3: Dataset summary and categorical data before treatment



```
── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────
   skim_variable                                 n_missing complete_rate        mean          sd        p0
 1 Administered_Daily                                    0         1         30473.     166946.   -1593072
 2 Administered_Cumulative                               0         1      14438588.   61681235.          0
 3 Administered_7_Day_Rolling_Average                 2820         0.968     29505.     152810.    -138218
 4 Admin_Dose_1_Daily                                    0         1         12177.     131625.   -2468411
 5 Admin_Dose_1_Cumulative                               0         1       6734796.   27972640.          0
 6 Admin_Dose_1_Day_Rolling_Average                   2820         0.968     12320.      86767.    -326573
 7 Administered_Dose1_Pop_Pct                            0         1            60.1        25.6         0
 8 Administered_daily_change_report                  21060         0.762     17755.     129866.          0
 9 Administered_daily_change_report_7dayroll         22140         0.75      35706.     171723.    -138218
10 Series_Complete_Daily                                 0         1         10395.     120321.    -523379
11 Series_Complete_Cumulative                            0         1       5666048.   23868426.          0
12 Series_Complete_Day_Rolling_Average                2820         0.968     10532.      78410.     -71931
13 Series_Complete_Pop_Pct                               0         1            50.7        24.1         0
14 Booster_Daily                                         0         1          5342.      46134.    -751692
15 Booster_Cumulative                                    0         1       1840204.    9765986.          0
16 Booster_7_Day_Rolling_Average                      2820         0.968      5193.      41199.      -2097
17 Additional_Doses_Vax_Pct                              0         1            24.9        23.5         0
18 Second_Booster_50Plus_Daily                           0         1          1658.      43769.     -40176
19 Second_Booster_50Plus_Cumulative                      0         1        281855.    2062760.          0
20 Second_Booster_50Plus_7_Day_Rolling_Average        2820         0.968      1222.      18472.       -170
21 Second_Booster_50Plus_Vax_Pct                         0         1            12.1        19.3         0
22 Bivalent_Booster_Daily                                0         1          2548.      65505.     -78273
23 Bivalent_Booster_Cumulative                           0         1        263516.    2570871.          0
24 Bivalent_Booster_7_Day_Rolling_Average             2820         0.968      1314.      16937.          0
25 Bivalent_Booster_Pop_Pct                              0         1             2.37        5.93        0
```

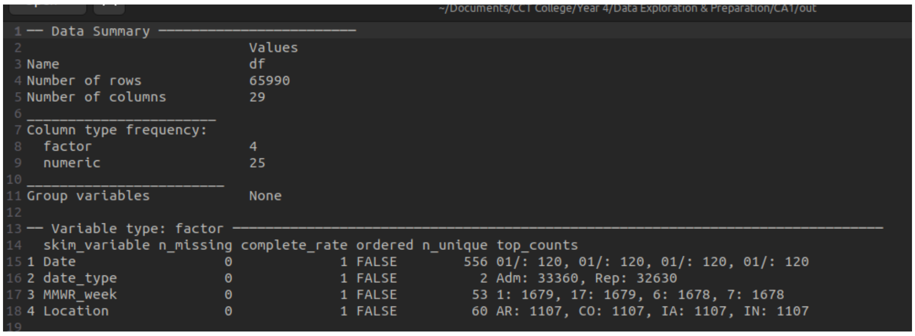
Figure 4: Numeric data before treatment



```
25 Bivalent_Booster_Pop_Pct
        p25       p50       p75        p100 hist
 1     741      4523     16276.     5610894 ▁▇▁▁▁
 2 986492.   3594100.  9668021   676728782 ▇▁▁▁▁
 3    1296      5430     16489     3508126 ▇▁▁▁▁
 4     147       954.     4974    30150525 ▇▁▁▁▁
 5 508493    1836880.  4330450   270227181 ▇▁▁▁▁
 6     199      1135      5381.    5011136 ▇▁▁▁▁
 7      48.2      65.2      78         100 ▁▁▂▇▅
 8       0         1      6765.    4581777 ▇▁▁▁▁
 9    2050      7791     20567     3508126 ▇▁▁▁▁
10      82       661      3835.   28709428 ▇▁▁▁▁
11 378088.   1512615   3837964.  230637348 ▇▁▁▁▁
12     128       790      4235     4837589 ▇▁▁▁▁
13      39.7      56.4      67.2      91.3 ▂▁▅▇▂
14       0        64      1171     2748886 ▁▇▁▁▁
15       0    131998.  1082856.  118438297 ▇▁▁▁▁
16       0       101      1273     1257266 ▇▁▁▁▁
17       0        26.9      47.4      67.7 ▇▁▂▅▂
18       0         0       139    11835302 ▇▁▁▁▁
19       0         0     72301    36744717 ▇▁▁▁▁
20       0         0       138     1855467 ▇▁▁▁▁
21       0         0        23.9      67.8 ▇▁▁▁▁
22       0         0         0    14759770 ▇▁▁▁▁
23       0         0         0    56478510 ▇▁▁▁▁
24       0         0         0      615808 ▇▁▁▁▁
25       0         0         0        34.3 ▇▁▁▁▁
```

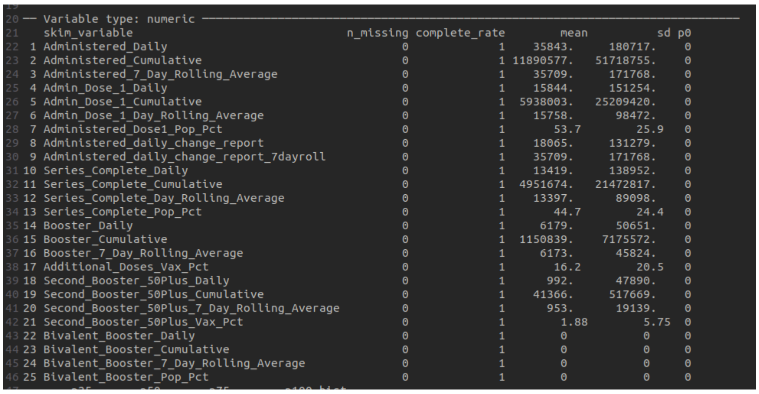
Figure 5: Numeric data before treatment (cont.)

Figures 6, 7 and 8 show the details after treatment:



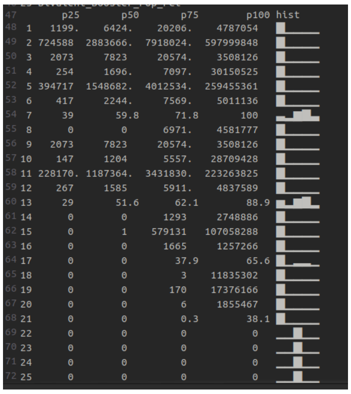
```
~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/out
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             65990
Number of columns          29
_______________________
Column type frequency:
  factor                   4
  numeric                  25
________________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 Date                  0             1 FALSE        556 01/: 120, 01/: 120, 01/: 120, 01/: 120
2 date_type             0             1 FALSE          2 Adm: 33360, Rep: 32630
3 MMWR_week             0             1 FALSE         53 1: 1679, 17: 1679, 6: 1678, 7: 1678
4 Location              0             1 FALSE         60 AR: 1107, CO: 1107, IA: 1107, IN: 1107
```

Figure 6: Dataset summary and categorical data after treatment



```
── Variable type: numeric ──────────────────────────────────────────────────────
   skim_variable                               n_missing complete_rate      mean        sd p0
 1 Administered_Daily                                  0             1    35843.   180717.   0
 2 Administered_Cumulative                             0             1 11890577. 51718755.   0
 3 Administered_7_Day_Rolling_Average                  0             1    35709.   171768.   0
 4 Admin_Dose_1_Daily                                  0             1    15844.   151254.   0
 5 Admin_Dose_1_Cumulative                             0             1  5938003. 25209420.   0
 6 Admin_Dose_1_Day_Rolling_Average                    0             1    15758.    98472.   0
 7 Administered_Dose1_Pop_Pct                          0             1       53.7      25.9  0
 8 Administered_daily_change_report                    0             1    18065.   131279.   0
 9 Administered_daily_change_report_7dayroll           0             1    35709.   171768.   0
10 Series_Complete_Daily                               0             1    13419.   138952.   0
11 Series_Complete_Cumulative                          0             1  4951674. 21472817.   0
12 Series_Complete_Day_Rolling_Average                 0             1    13397.    89098.   0
13 Series_Complete_Pop_Pct                             0             1       44.7      24.4  0
14 Booster_Daily                                       0             1     6179.    50651.   0
15 Booster_Cumulative                                  0             1  1150839.  7175572.   0
16 Booster_7_Day_Rolling_Average                       0             1     6173.    45824.   0
17 Additional_Doses_Vax_Pct                            0             1       16.2      20.5  0
18 Second_Booster_50Plus_Daily                         0             1      992.    47890.   0
19 Second_Booster_50Plus_Cumulative                    0             1    41366.   517669.   0
20 Second_Booster_50Plus_7_Day_Rolling_Average         0             1      953.    19139.   0
21 Second_Booster_50Plus_Vax_Pct                       0             1        1.88      5.75 0
22 Bivalent_Booster_Daily                              0             1        0         0    0
23 Bivalent_Booster_Cumulative                         0             1        0         0    0
24 Bivalent_Booster_7_Day_Rolling_Average              0             1        0         0    0
25 Bivalent_Booster_Pop_Pct                            0             1        0         0    0
```

Figure 7: Numeric data after treatment



```
        p25       p50       p75      p100 hist
 1    1199.     6424.    20206.   4787054 ▇▁▁▁▁
 2 724588   2883666.  7918024. 597999848 ▇▁▁▁▁
 3    2073      7823     20574.   3508126 ▇▁▁▁▁
 4     254      1696.     7097.  30150525 ▇▁▁▁▁
 5 394717   1548682.  4012534. 259455361 ▇▁▁▁▁
 6     417      2244.     7569.   5011136 ▇▁▁▁▁
 7      39        59.8      71.8      100 ▂▁▆▇▂
 8       0         0      6971.   4581777 ▇▁▁▁▁
 9    2073      7823     20574.   3508126 ▇▁▁▁▁
10     147      1204      5557.  28709428 ▇▁▁▁▁
11 228170.  1187364.  3431830. 223263825 ▇▁▁▁▁
12     267      1585      5911.   4837589 ▇▁▁▁▁
13      29        51.6      62.1       88.9 ▃▁▆▇▂
14       0         0      1293    2748886 ▇▁▁▁▁
15       0         1    579131  107058288 ▇▁▁▁▁
16       0         0      1665    1257266 ▇▁▁▁▁
17       0         0        37.9       65.6 ▇▁▂▂▁
18       0         0         3   11835302 ▇▁▁▁▁
19       0         0       170   17376166 ▇▁▁▁▁
20       0         0         6    1855467 ▇▁▁▁▁
21       0         0         0.3       38.1 ▇▁▁▁▁
22       0         0         0          0 ▁▁▇▁▁
23       0         0         0          0 ▁▁▇▁▁
24       0         0         0          0 ▁▁▇▁▁
25       0         0         0          0 ▁▁▇▁▁
```

Figure 8: Numeric data after treatment (cont.)

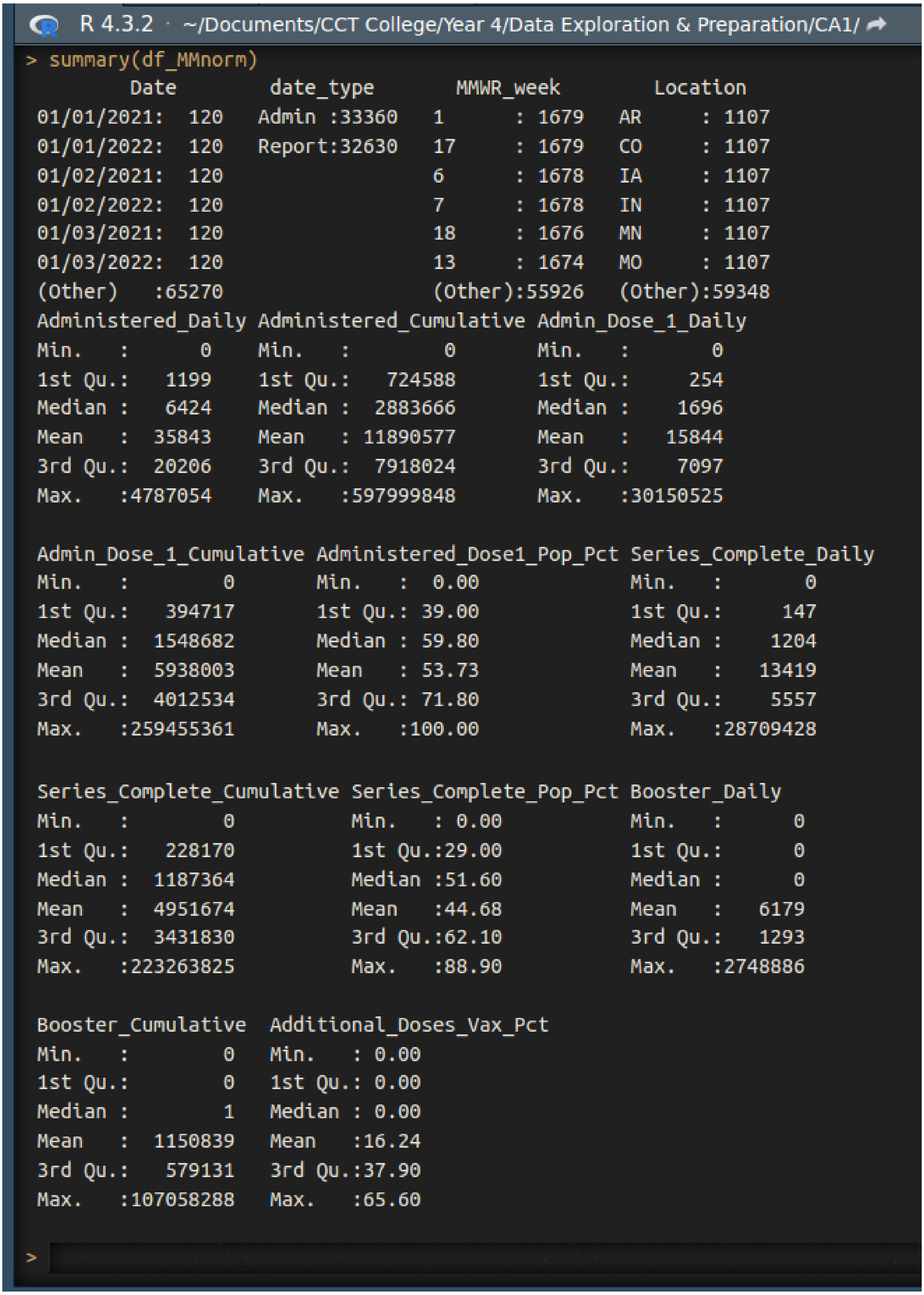
Observations with missing or negative values were dropped, which reduced the number of rows in the dataset to 65,990.

Unwanted columns were also dropped, reducing the total number to 15. The following are the columns that were kept in the dataset:

- Date
- date_type
- MMWR_week
- Location
- Administered_Daily
- Administered_Cumulative
- Admin_Dose_1_Daily
- Admin_Dose_1_Cumulative
- Administered_Dose1_Pop_Pct
- Series_Complete_Daily
- Series_Complete_Cumulative
- Series_Complete_Pop_Pct
- Booster_Daily
- Booster_Cumulative
- Additional_Doses_Vax_Pct

Feature scaling

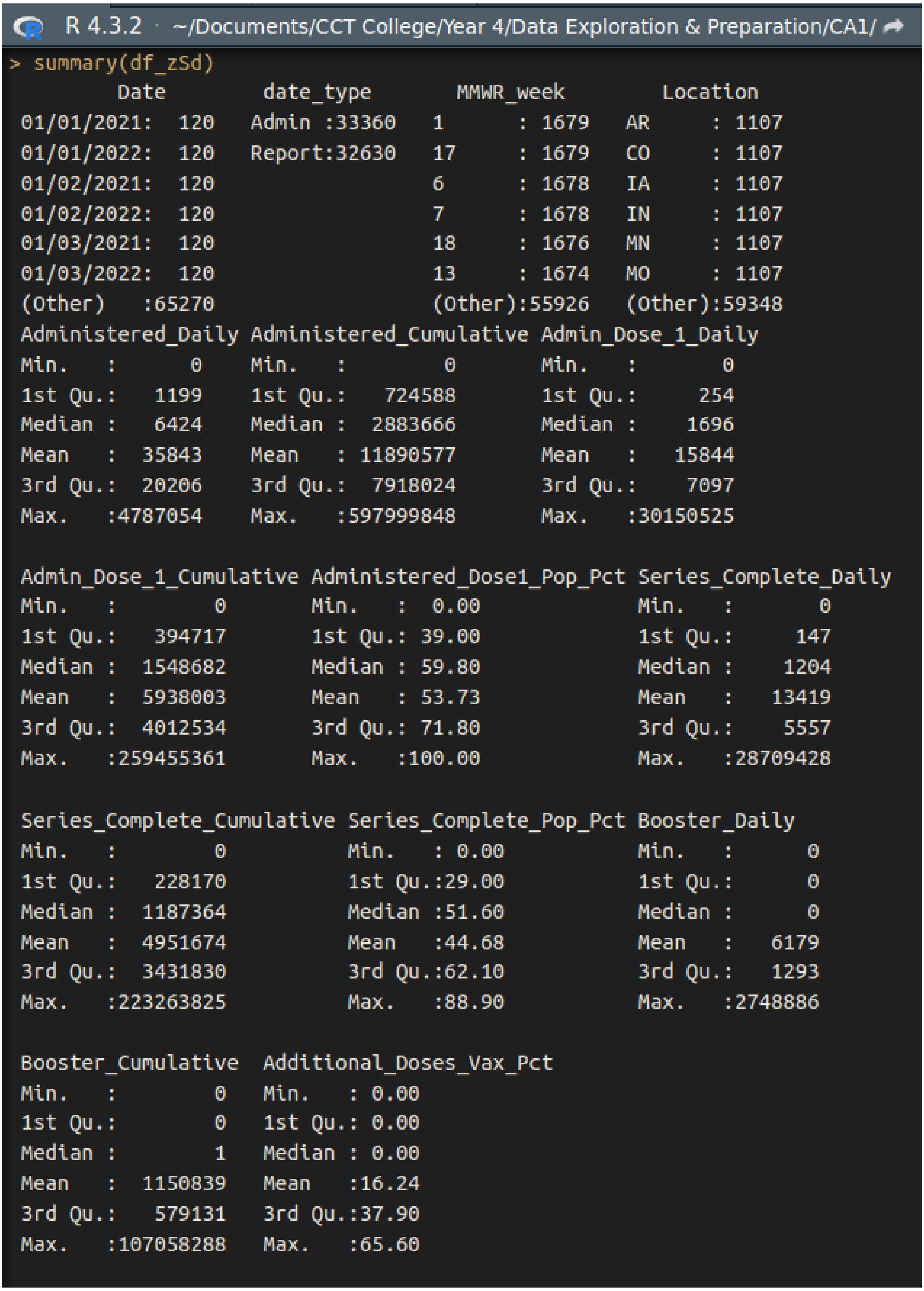
Min-Max Normalization, Z-Score Standardization and Robust Scaler were applied to all numeric values of the dataset, Figures 9, 10 and 11, below, show the results of each scaling method:



R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

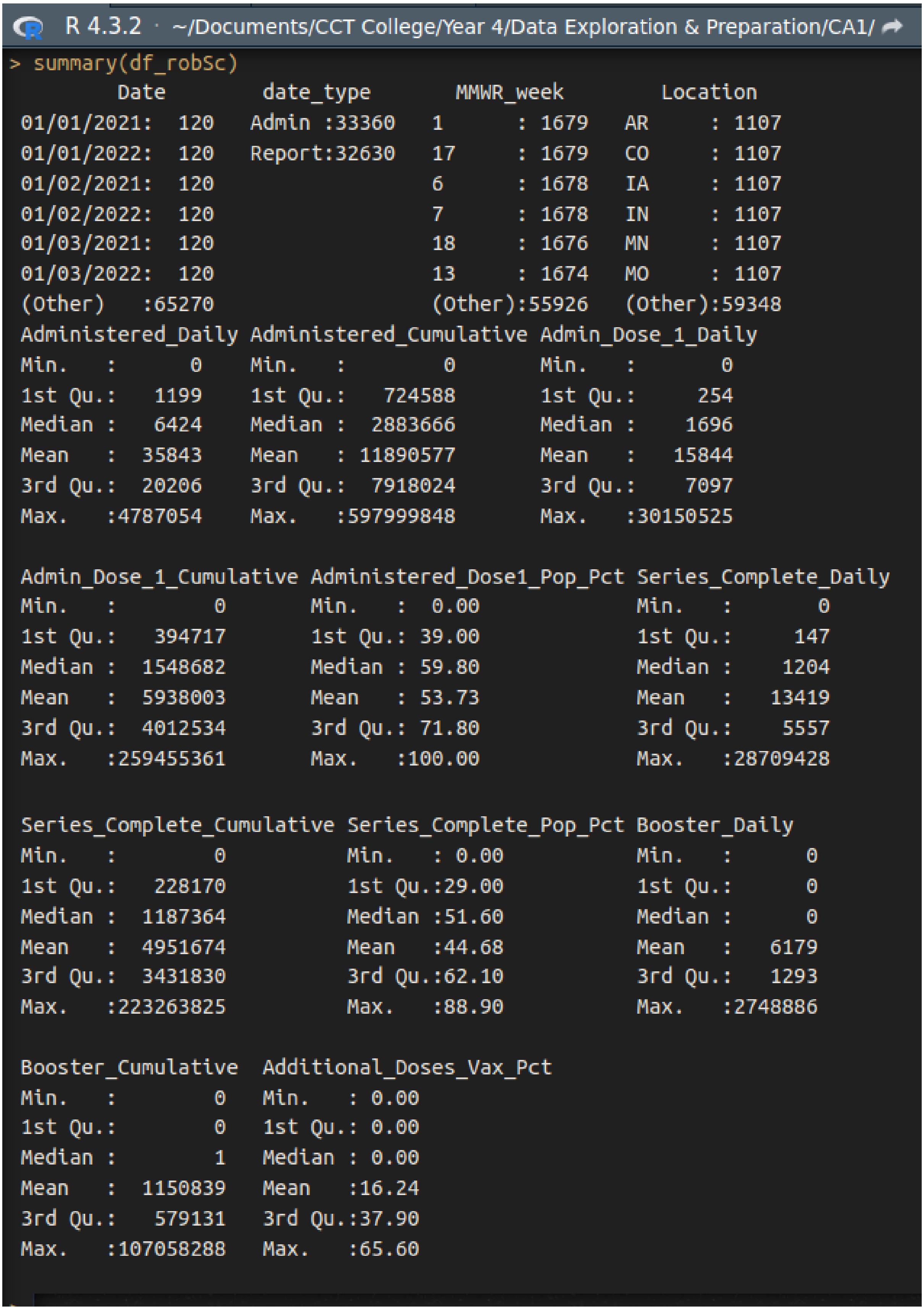
```
> summary(df_MMnorm)
         Date            date_type         MMWR_week         Location    
 01/01/2021:  120   Admin :33360   1       : 1679   AR      : 1107  
 01/01/2022:  120   Report:32630   17      : 1679   CO      : 1107  
 01/02/2021:  120                  6       : 1678   IA      : 1107  
 01/02/2022:  120                  7       : 1678   IN      : 1107  
 01/03/2021:  120                  18      : 1676   MN      : 1107  
 01/03/2022:  120                  13      : 1674   MO      : 1107  
 (Other)   :65270                  (Other):55926   (Other):59348  
 Administered_Daily Administered_Cumulative Admin_Dose_1_Daily
 Min.   :      0    Min.   :        0        Min.   :       0  
 1st Qu.:   1199    1st Qu.:   724588        1st Qu.:     254  
 Median :   6424    Median :  2883666        Median :    1696  
 Mean   :  35843    Mean   : 11890577        Mean   :   15844  
 3rd Qu.:  20206    3rd Qu.:  7918024        3rd Qu.:    7097  
 Max.   :4787054    Max.   :597999848        Max.   :30150525  
                                                               
 Admin_Dose_1_Cumulative Administered_Dose1_Pop_Pct Series_Complete_Daily
 Min.   :        0       Min.   :  0.00             Min.   :       0     
 1st Qu.:   394717       1st Qu.: 39.00             1st Qu.:     147     
 Median :  1548682       Median : 59.80             Median :    1204     
 Mean   :  5938003       Mean   : 53.73             Mean   :   13419     
 3rd Qu.:  4012534       3rd Qu.: 71.80             3rd Qu.:    5557     
 Max.   :259455361       Max.   :100.00             Max.   :28709428     
                                                                         
 Series_Complete_Cumulative Series_Complete_Pop_Pct Booster_Daily    
 Min.   :        0          Min.   : 0.00           Min.   :      0  
 1st Qu.:   228170          1st Qu.:29.00           1st Qu.:      0  
 Median :  1187364          Median :51.60           Median :      0  
 Mean   :  4951674          Mean   :44.68           Mean   :   6179  
 3rd Qu.:  3431830          3rd Qu.:62.10           3rd Qu.:   1293  
 Max.   :223263825          Max.   :88.90           Max.   :2748886  
                                                                    
 Booster_Cumulative  Additional_Doses_Vax_Pct
 Min.   :        0   Min.   : 0.00           
 1st Qu.:        0   1st Qu.: 0.00           
 Median :        1   Median : 0.00           
 Mean   :  1150839   Mean   :16.24           
 3rd Qu.:   579131   3rd Qu.:37.90           
 Max.   :107058288   Max.   :65.60           

> 
```

Figure 9: Min-Max normalized dataset


R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

```
> summary(df_zSd)
         Date        date_type       MMWR_week         Location    
 01/01/2021:  120   Admin :33360   1      : 1679   AR     : 1107  
 01/01/2022:  120   Report:32630   17     : 1679   CO     : 1107  
 01/02/2021:  120                  6      : 1678   IA     : 1107  
 01/02/2022:  120                  7      : 1678   IN     : 1107  
 01/03/2021:  120                  18     : 1676   MN     : 1107  
 01/03/2022:  120                  13     : 1674   MO     : 1107  
 (Other)   :65270                  (Other):55926   (Other):59348  
 Administered_Daily Administered_Cumulative Admin_Dose_1_Daily
 Min.   :      0    Min.   :        0       Min.   :       0  
 1st Qu.:   1199    1st Qu.:   724588       1st Qu.:     254  
 Median :   6424    Median :  2883666       Median :    1696  
 Mean   :  35843    Mean   : 11890577       Mean   :   15844  
 3rd Qu.:  20206    3rd Qu.:  7918024       3rd Qu.:    7097  
 Max.   :4787054    Max.   :597999848       Max.   :30150525  
                                                              
 Admin_Dose_1_Cumulative Administered_Dose1_Pop_Pct Series_Complete_Daily
 Min.   :        0       Min.   :  0.00             Min.   :       0     
 1st Qu.:   394717       1st Qu.: 39.00             1st Qu.:     147     
 Median :  1548682       Median : 59.80             Median :    1204     
 Mean   :  5938003       Mean   : 53.73             Mean   :   13419     
 3rd Qu.:  4012534       3rd Qu.: 71.80             3rd Qu.:    5557     
 Max.   :259455361       Max.   :100.00             Max.   :28709428     
                                                                         
 Series_Complete_Cumulative Series_Complete_Pop_Pct Booster_Daily    
 Min.   :        0          Min.   : 0.00           Min.   :      0  
 1st Qu.:   228170          1st Qu.:29.00           1st Qu.:      0  
 Median :  1187364          Median :51.60           Median :      0  
 Mean   :  4951674          Mean   :44.68           Mean   :   6179  
 3rd Qu.:  3431830          3rd Qu.:62.10           3rd Qu.:   1293  
 Max.   :223263825          Max.   :88.90           Max.   :2748886  
                                                                     
 Booster_Cumulative  Additional_Doses_Vax_Pct
 Min.   :        0   Min.   : 0.00           
 1st Qu.:        0   1st Qu.: 0.00           
 Median :        1   Median : 0.00           
 Mean   :  1150839   Mean   :16.24           
 3rd Qu.:   579131   3rd Qu.:37.90           
 Max.   :107058288   Max.   :65.60           
```

Figure 10: Z-Score standardized dataset


R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

```
> summary(df_robSc)
         Date        date_type        MMWR_week         Location    
 01/01/2021:  120   Admin :33360   1      : 1679   AR     : 1107  
 01/01/2022:  120   Report:32630   17     : 1679   CO     : 1107  
 01/02/2021:  120                  6      : 1678   IA     : 1107  
 01/02/2022:  120                  7      : 1678   IN     : 1107  
 01/03/2021:  120                  18     : 1676   MN     : 1107  
 01/03/2022:  120                  13     : 1674   MO     : 1107  
 (Other)   :65270                  (Other):55926   (Other):59348  
 Administered_Daily Administered_Cumulative Admin_Dose_1_Daily
 Min.   :      0    Min.   :        0       Min.   :       0  
 1st Qu.:   1199    1st Qu.:   724588       1st Qu.:     254  
 Median :   6424    Median :  2883666       Median :    1696  
 Mean   :  35843    Mean   : 11890577       Mean   :   15844  
 3rd Qu.:  20206    3rd Qu.:  7918024       3rd Qu.:    7097  
 Max.   :4787054    Max.   :597999848       Max.   :30150525  
                                                              
 Admin_Dose_1_Cumulative Administered_Dose1_Pop_Pct Series_Complete_Daily
 Min.   :        0       Min.   :  0.00             Min.   :       0     
 1st Qu.:   394717       1st Qu.: 39.00             1st Qu.:     147     
 Median :  1548682       Median : 59.80             Median :    1204     
 Mean   :  5938003       Mean   : 53.73             Mean   :   13419     
 3rd Qu.:  4012534       3rd Qu.: 71.80             3rd Qu.:    5557     
 Max.   :259455361       Max.   :100.00             Max.   :28709428     
                                                                         
 Series_Complete_Cumulative Series_Complete_Pop_Pct Booster_Daily    
 Min.   :        0          Min.   : 0.00           Min.   :      0  
 1st Qu.:   228170          1st Qu.:29.00           1st Qu.:      0  
 Median :  1187364          Median :51.60           Median :      0  
 Mean   :  4951674          Mean   :44.68           Mean   :   6179  
 3rd Qu.:  3431830          3rd Qu.:62.10           3rd Qu.:   1293  
 Max.   :223263825          Max.   :88.90           Max.   :2748886  
                                                                     
 Booster_Cumulative  Additional_Doses_Vax_Pct
 Min.   :        0   Min.   : 0.00           
 1st Qu.:        0   1st Qu.: 0.00           
 Median :        1   Median : 0.00           
 Mean   :  1150839   Mean   :16.24           
 3rd Qu.:   579131   3rd Qu.:37.90           
 Max.   :107058288   Max.   :65.60           
```

Figure 11: Robust scaled dataset

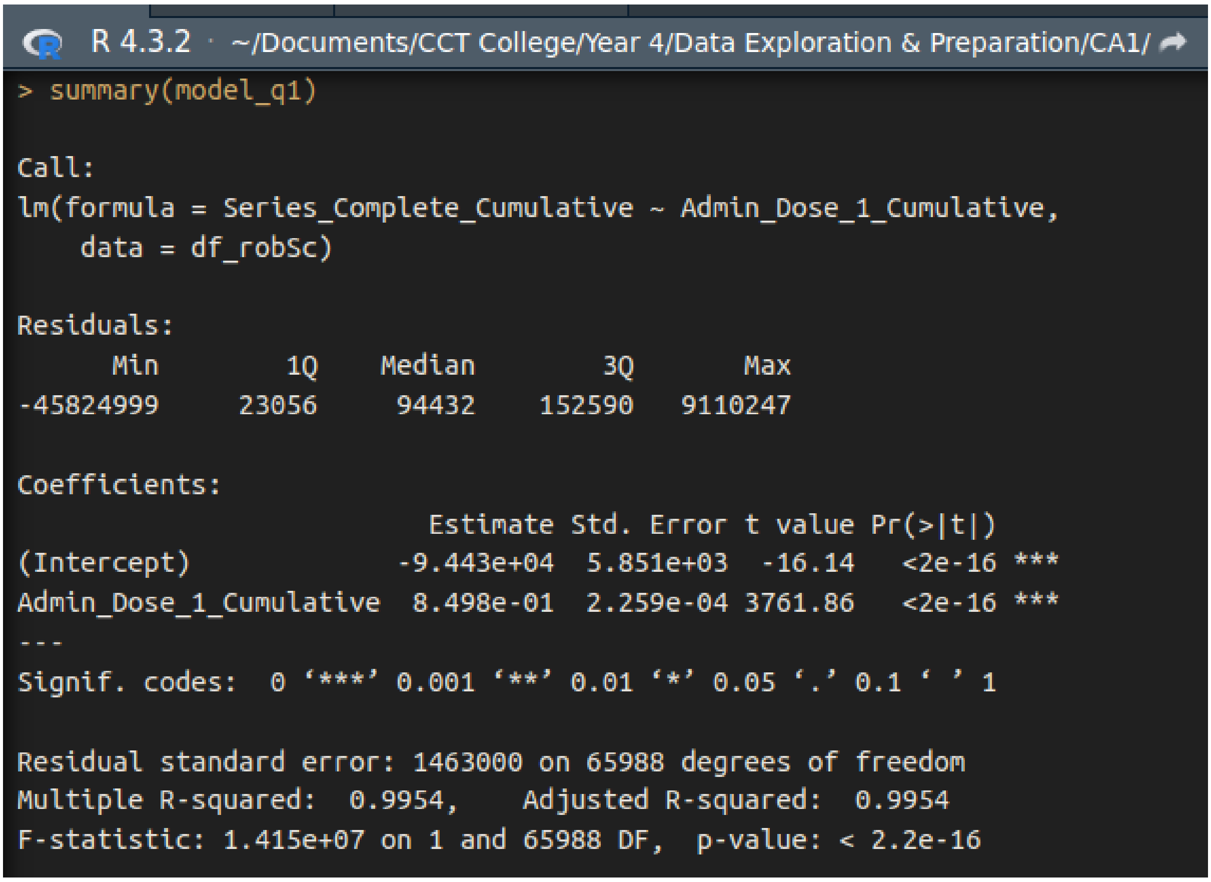
Exploratory Data Analysis

For the EDA part of this assignment, it was attempted to answer the following questions:

1. “Are people who took the 1st dose more likely to complete the series?”
2. "Are people who completed the series more likely to take the booster?"

Feature correlation

Figure 12 shows the correlation score calculated through linear regression for the variables Admin_Dose_1_Cumulative (independent) and Series_Complete_Cumulative (dependent), aimed at Question 1:

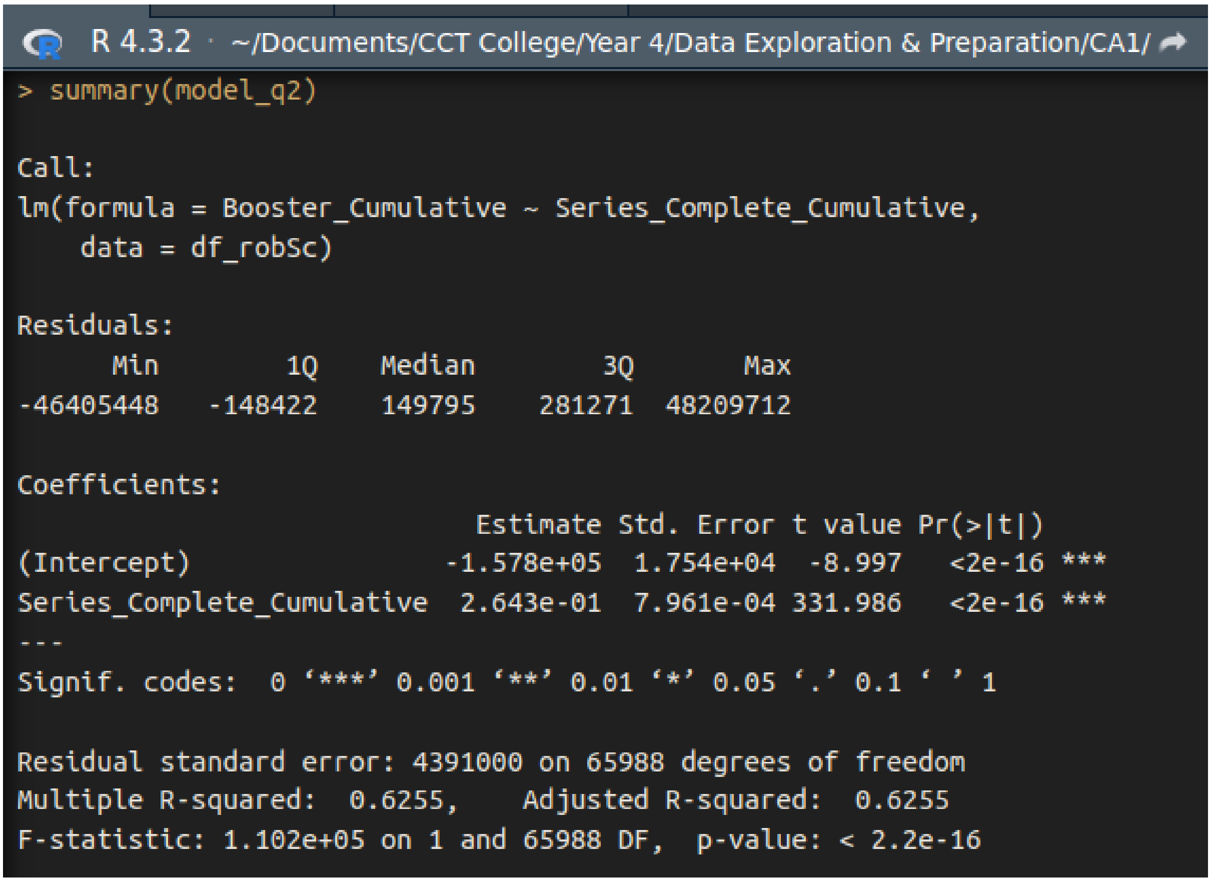

R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

```
> summary(model_q1)

Call:
lm(formula = Series_Complete_Cumulative ~ Admin_Dose_1_Cumulative,
    data = df_robSc)

Residuals:
      Min        1Q    Median        3Q       Max
-45824999     23056     94432    152590   9110247

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -9.443e+04  5.851e+03  -16.14   <2e-16 ***
Admin_Dose_1_Cumulative  8.498e-01  2.259e-04 3761.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1463000 on 65988 degrees of freedom
Multiple R-squared:  0.9954,    Adjusted R-squared:  0.9954
F-statistic: 1.415e+07 on 1 and 65988 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of linear regression model for Question 1

The Adjusted R-Squared of 0.9954 suggests a strong correlation between the two variables.

Figure 13 shows the correlation score calculated through linear regression for the variables Series_Complete_Cumulative (independent) and Booster_Cumulative (dependent), aimed at Question 2:



R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

```
> summary(model_q2)

Call:
lm(formula = Booster_Cumulative ~ Series_Complete_Cumulative,
    data = df_robSc)

Residuals:
      Min        1Q    Median        3Q       Max
-46405448   -148422    149795    281271  48209712

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.578e+05  1.754e+04  -8.997   <2e-16 ***
Series_Complete_Cumulative  2.643e-01  7.961e-04 331.986   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4391000 on 65988 degrees of freedom
Multiple R-squared:  0.6255,	Adjusted R-squared:  0.6255
F-statistic: 1.102e+05 on 1 and 65988 DF,  p-value: < 2.2e-16
```

Figure 13: Summary of linear regression model for Question 2

The Adjusted R-Squared of 0.6255 suggests a mild correlation between the two variables.

Data exploration

Figures 14 and 15, below, explore the aforementioned correlations graphically:

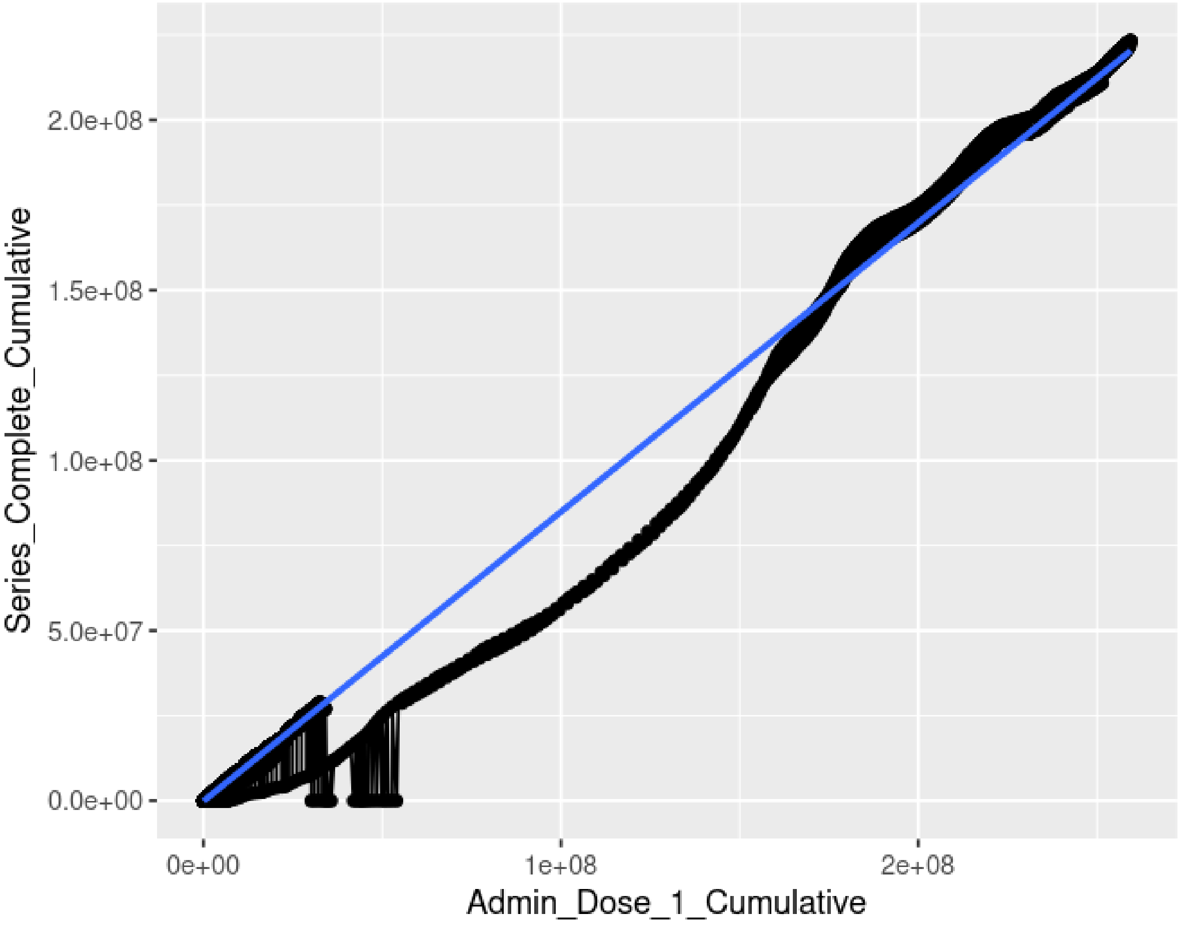


Figure 14: Accumulated Number of Series Complete vs Accumulated Number of Administered Dose 1s

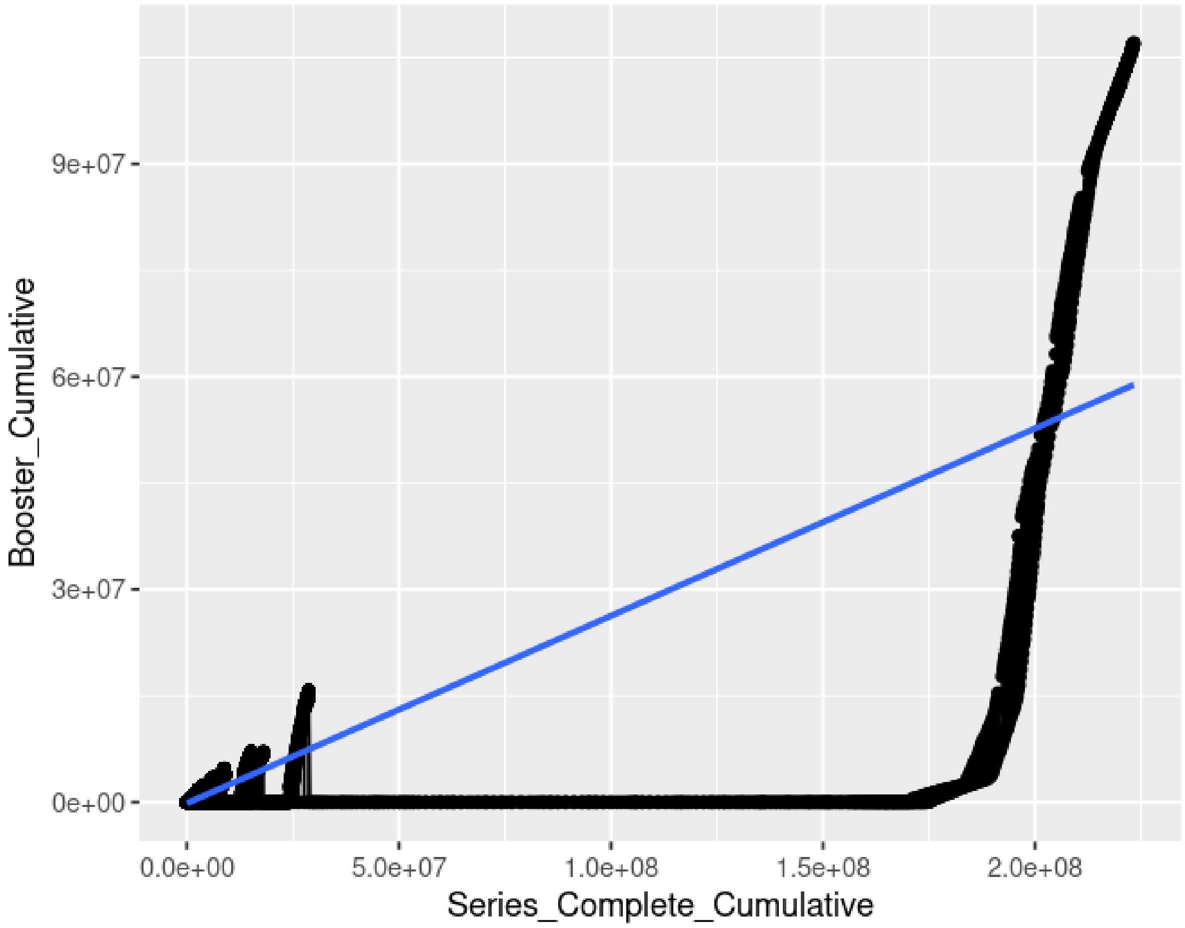


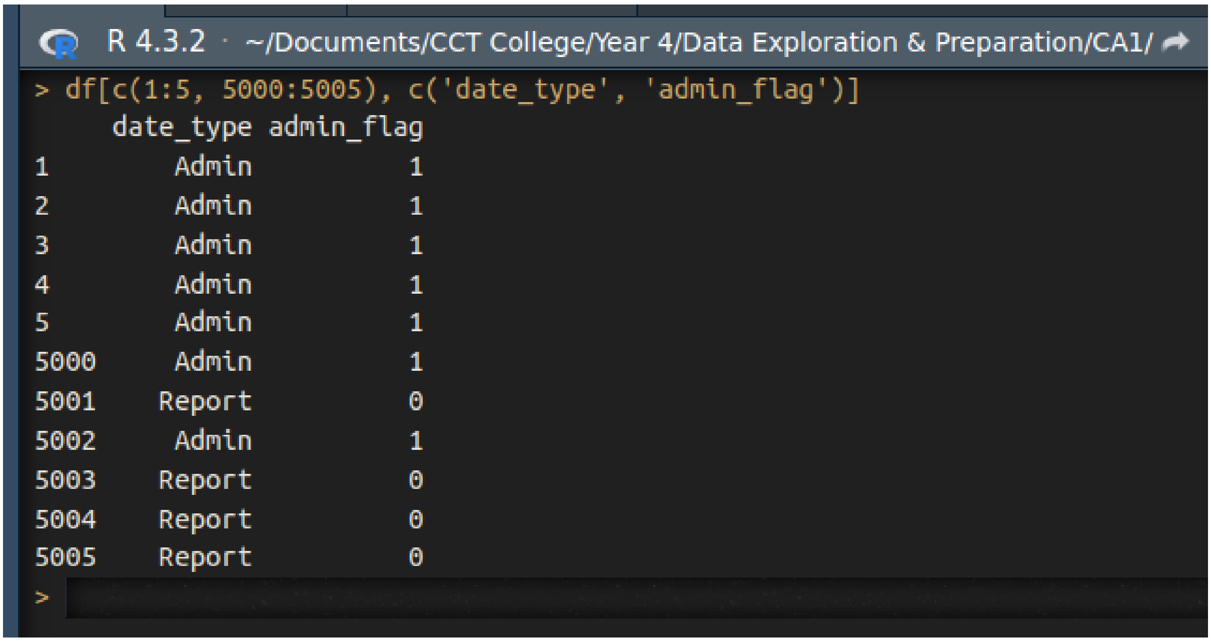
Figure 15: Accumulated Number of Booster Doses vs Accumulated Number of Series Complete

Principal Component Analysis

This section is about applying PCA to achieve dimensionality reduction and make data analysis both cheaper and more robust.

Dummy encoding

For the dummy encoding, the label admin-flag was added, which assumes the value 0 if the variable data-type is equal to Report and 1 if equal to Admin:

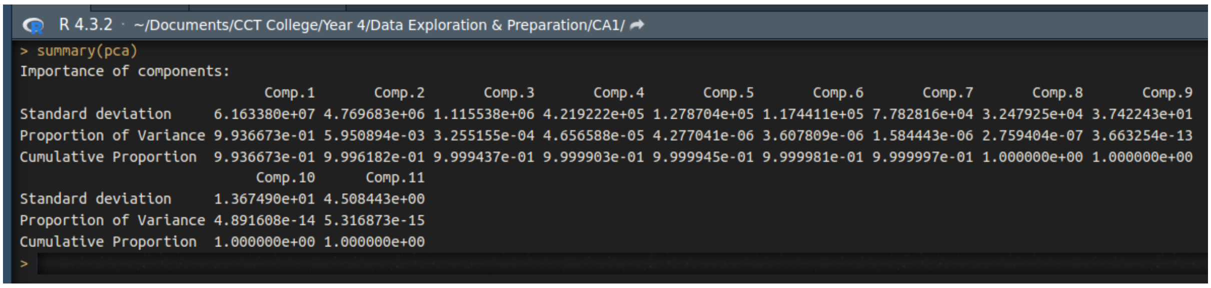

R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

```
> df[c(1:5, 5000:5005), c('date_type', 'admin_flag')]
     date_type admin_flag
1        Admin          1
2        Admin          1
3        Admin          1
4        Admin          1
5        Admin          1
5000     Admin          1
5001    Report          0
5002     Admin          1
5003    Report          0
5004    Report          0
5005    Report          0
>
```

Figure 16: Dummy encoding admin_flag

Component profile

Figure 17 shows each variable in the dataset as a profiled component:


R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

```
> summary(pca)
Importance of components:
                             Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6       Comp.7       Comp.8       Comp.9
Standard deviation     6.163380e+07 4.769683e+06 1.115538e+06 4.219222e+05 1.278704e+05 1.174411e+05 7.782816e+04 3.247925e+04 3.742243e+01
Proportion of Variance 9.936673e-01 5.950894e-03 3.255155e-04 4.656588e-05 4.277041e-06 3.607809e-06 1.584443e-06 2.759404e-07 3.663254e-13
Cumulative Proportion  9.936673e-01 9.996182e-01 9.999437e-01 9.999903e-01 9.999945e-01 9.999981e-01 9.999997e-01 1.000000e+00 1.000000e+00
                            Comp.10      Comp.11
Standard deviation     1.367490e+01 4.508443e+00
Proportion of Variance 4.891608e-14 5.316873e-15
Cumulative Proportion  1.000000e+00 1.000000e+00
>
```

Figure 17: Component profiles

Figure 18 shows a bar chart where the components are sorted by their variance, suggesting the impact they each have on the analysis:

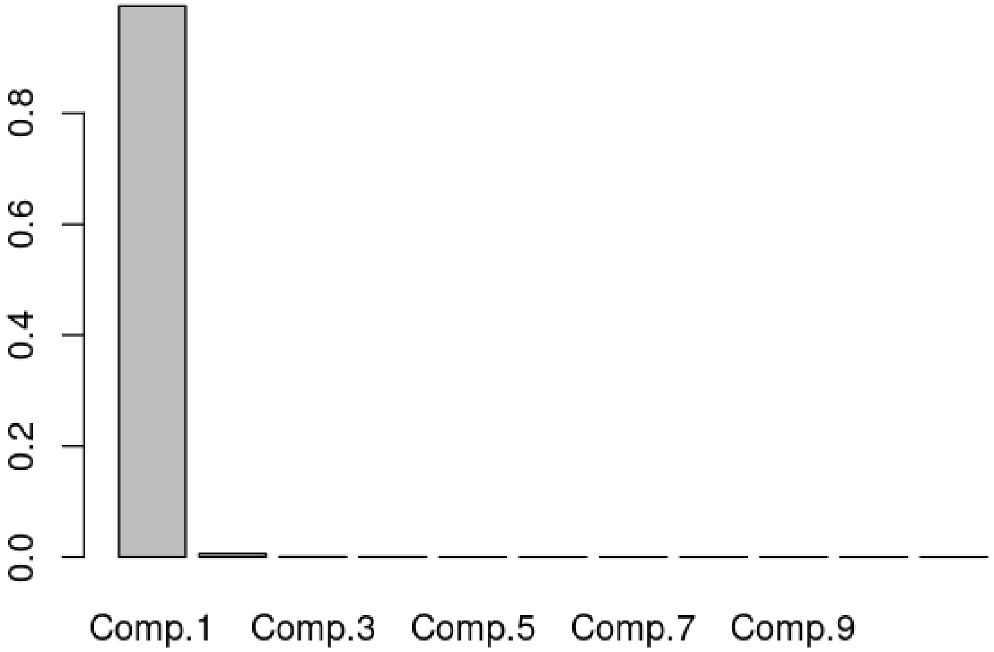
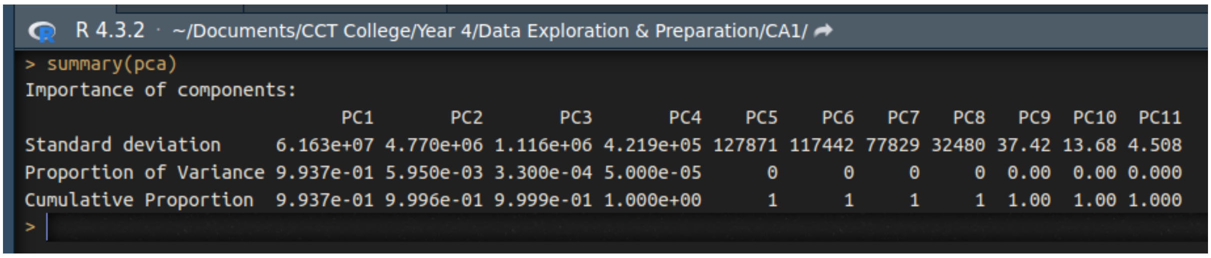


Figure 18: Components by variance

Dimensionality reduction

In Figure 19, it can be seen that the components are reduced in terms of their relevance to the dataset:


R 4.3.2 · ~/Documents/CCT College/Year 4/Data Exploration & Preparation/CA1/

```
> summary(pca)
Importance of components:
                             PC1       PC2       PC3       PC4    PC5    PC6   PC7   PC8   PC9  PC10  PC11
Standard deviation     6.163e+07 4.770e+06 1.116e+06 4.219e+05 127871 117442 77829 32480 37.42 13.68 4.508
Proportion of Variance 9.937e-01 5.950e-03 3.300e-04 5.000e-05      0      0     0     0  0.00  0.00 0.000
Cumulative Proportion  9.937e-01 9.996e-01 9.999e-01 1.000e+00      1      1     1     1  1.00  1.00 1.000
> 
```

Figure 19: Principal Components

Figure 20 shows the biplot of the two main principal components identified in the previous step, PC1 and PC2:

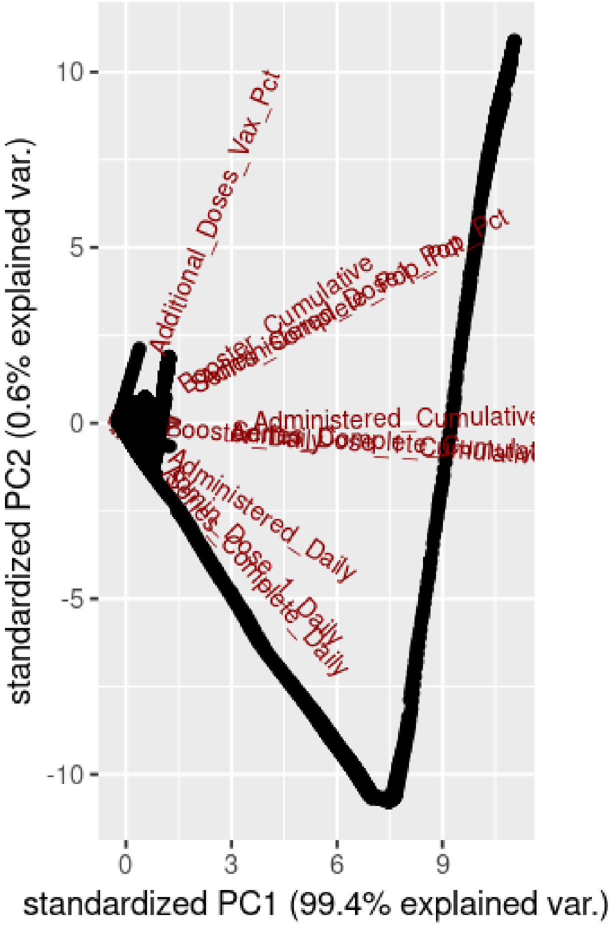


Figure 20: PC2 vs PC1

Source Code

The source code for the analysis above was written in the R language, as per the table below:

```
# DEP_Lv8_CA1
#
# CCT College Dublin
# Bachelor of Science Honours in Computing in Information Technology
# Data Exploration & Preparation - Y4M3
# Year 4, Semester 7
# Continuous Assessment 1
#
# Lecturer name: Dr. Muhammad Iqbal
# Lecturer email: miqbal@cct.ie
#
# Student Name: Mateus Fonseca Campos
# Student Number: 2023327
# Student Email: 2023327@student.cct.ie
#
# Submission date: 3 December 2023
#
# GitHub: https://github.com/2023327cctcollege/DEP_Lv8_CA1
# ____


# installing all the necessary packages at once
packages <- c('tidyr', 'dplyr', 'skimr', 'ggplot2', 'devtools')
for (p in packages) {
  if (!(p %in% rownames(installed.packages()))) {
    install.packages(p, character.only = TRUE)
  }
  library(p, character.only = TRUE)
}

# installing ggbiplot from GitHub repository
# docs suggests that ggbiplot be loaded before dplyr
# unload dplyr, load ggbiplot, then load dplyr again
install_github("vqv/ggbiplot")
unload('dplyr')
library(ggbiplot)
library(dplyr)

# read dataset from CSV file
# available at
https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3
yt2
# accessed on 3 December 2023
df <-
read.csv('COVID-19_Vaccination_Trends_in_the_United_States_National_and_Jurisdictional.csv'
, stringsAsFactors = TRUE)
```

```
# 1. Data Preparation

# 1.1. Variable classification

# quick summary of variables in the dataset
# categorical, discrete or continuous
var_class <- function(df) {
  factors <- df %>% select_if(is.factor)
  numerics <- df %>% select_if(is.numeric)
  discretes <- numerics %>% select_if(function(x) all(x %% 1 == 0 | is.na(x)))
  continuous <- numerics %>% select_if(Negate(function(x) all(x %% 1 == 0 | is.na(x))))

  cat(sprintf('Categorical [%d]:\n', ncol(factors)))
  for (col in colnames(factors)) {
    cat(sprintf('\t%s\n', col))
  }

  cat(sprintf('\nDiscrete [%d]:\n', ncol(discretes)))
  for (col in colnames(discretes)) {
    cat(sprintf('\t%s\n', col))
  }

  cat(sprintf('\nContinuous [%d]:\n', ncol(continuous)))
  for (col in colnames(continuous)) {
    cat(sprintf('\t%s\n', col))
  }
}

# write var_class to file before change
sink('./out/fig_2.txt')
var_class(df)
sink()

# make MMWR_week a factor
# variable is numeric however it labels the week number
# to be used for categorization rather than calculation
df$MMWR_week <- as.factor(df$MMWR_week)

# write var_class to file after change
sink('./out/fig_3.txt')
var_class(df)
sink()

# write skim to file before change
sink('./out/fig_4-6.txt')
skim(df)
sink()

# replace negative values with NA
# drop rows that contain NA
df[df < 0] <- NA
df <- drop_na(df)

# write skim to file after change
sink('./out/fig_7-9.txt')
skim(df)
sink()

# keep only desired columns, drop the rest
df <- df[, c(1:6, 8, 9, 11, 14, 15, 17:19, 21)]


# 1.2. Feature scaling

# min-max normalization function
MMnorm <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

# z-score standardization function
zSd <- function(x) {
  return((x - mean(x)) / (sd(x)))
}

# robust scaler scaling function
robSc <- function(x) {
  return((x - median(x)) / (quantile(x, 0.75) - quantile(x, 0.25)))
}

# normalized/standardized/scaled versions of dataframe
df_MMnorm <- df
df_zSd <- df
df_robSc <- df

# apply scaling functions respectively
df_MMnorm[, 5:15] <- apply(df[, 5:15], 2, MMnorm)
df_zSd[, 5:15] <- apply(df[, 5:15], 2, zSd)
df_robSc[, 5:15] <- apply(df[, 5:15], 2, robSc)


# 2. Exploratory Data Analysis (EDA)

# Question 1: "Are people who took the 1st dose more likely to complete the series?"
# Question 2: "Are people who completed the series more likely to take the booster?"

# 2.1. Feature correlation

# Q1
# linear regression model
# correlation factor in summary
```

```
model_q1 <- lm(Series_Complete_Cumulative ~ Admin_Dose_1_Cumulative, df_robSc)

# Q2
# linear regression model
# correlation factor in summary
model_q2 <- lm(Booster_Cumulative ~ Series_Complete_Cumulative, df_robSc)


# 2.2. Data exploration

# Q1
# line + scatter plot with linear regression
ggplot(data = df_robSc, aes(x = Admin_Dose_1_Cumulative, y = Series_Complete_Cumulative)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = 'lm')

# Q2
# line + scatter plot with linear regression
ggplot(data = df_robSc, aes(x = Series_Complete_Cumulative, y = Booster_Cumulative)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = 'lm')


# 3. Principal Component Analysis (PCA)

# 3.1. Dummy encoding

# add dummy encoding flag for date_type
df_robSc <- transform(df, admin_flag=ifelse(date_type == 'Admin', 1, 0))


# 3.2. Component profile

# plotting a barplot of all components sorted by variance
pca <- princomp(df_robSc[, c(5:15)])
summary(pca)
pca$var$exp <- pca$sdev^2 / sum(pca$sdev^2)
barplot(pca$var$exp)


# 3.3. Dimensionality reduction

# plotting a biplot of PC1 against PC2 (the two main components)
pca <- prcomp(df_robSc[, c(5:15)], center = TRUE, scale. = FALSE)
summary(pca)
ggbiplot(pca)
```

Table 2: R script source code

The code is also available from the project's GitHub repository (Campos, 2023).

Conclusion

Didn’t really have time to do much at all. :\

References

Campos, M.F. (2023) *DEP_Lv8_CA1*. Available at: https://github.com/2023327cctcollege/DEP_Lv8_CA1 (Accessed 3 December 2023).

Centers for Disease Control and Prevention (2023) *COVID-19 Vaccination Trends in the United States,National and Jurisdictional*. Available at: https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2 (Accessed 3 December 2023).