DSC 680 Project 1

March 19, 2020

[1]:
```
import csv
import pandas as pd
import numpy as np

#Import file and display first 5 rows
fludf = pd.read_csv('C:/Users/Christine/Documents/Bellevue/DSC 680/Project 1/
 ↪fluprint_export.csv')
fludf.head()
```

[1]:
```
   donor_id  study_id  gender       race  visit_id  visit_year  visit_day  \
0       813        15  Female  Caucasian      2937        2014          0
1       813        15  Female  Caucasian      2937        2014          0
2       813        15  Female  Caucasian      2937        2014          0
3       813        15  Female  Caucasian      2937        2014          0
4       813        15  Female  Caucasian      2937        2014          0

  visit_type_hai  visit_age  cmv_status  …  vaccinated_2yr_prior  \
0            pre       23.0         0.0  …                   1.0
1            pre       23.0         0.0  …                   1.0
2            pre       23.0         0.0  …                   1.0
3            pre       23.0         0.0  …                   1.0
4            pre       23.0         0.0  …                   1.0

   vaccine_type_2yr_prior  vaccinated_3yr_prior  vaccine_type_3yr_prior  \
0                     2.0                   1.0                     2.0
1                     2.0                   1.0                     2.0
2                     2.0                   1.0                     2.0
3                     2.0                   1.0                     2.0
4                     2.0                   1.0                     2.0

   vaccinated_4yr_prior  vaccine_type_4yr_prior  vaccinated_5yr_prior  \
0                   1.0                     2.0                   1.0
1                   1.0                     2.0                   1.0
2                   1.0                     2.0                   1.0
3                   1.0                     2.0                   1.0
4                   1.0                     2.0                   1.0
```

```
   vaccine_type_5yr_prior  influenza_infection_history  \
0                     2.0                            0
1                     2.0                            0
2                     2.0                            0
3                     2.0                            0
4                     2.0                            0

  influenza_hospitalization
0                         0
1                         0
2                         0
3                         0
4                         0

[5 rows x 38 columns]
```

[2]:
```
#Display statistics for numerical variables
fludf.describe()
```

[2]:
```
            donor_id       study_id      visit_id     visit_year  \
count  156118.000000  156118.000000  156118.000000  156118.000000
mean      392.616181      19.359709     896.606452    2011.774241
std       190.507183       3.234831     679.654527       1.761023
min         1.000000      15.000000       1.000000    2007.000000
25%       240.000000      18.000000     396.000000    2011.000000
50%       402.000000      18.000000     871.000000    2013.000000
75%       501.000000      21.000000    1148.000000    2013.000000
max       813.000000      30.000000    2937.000000    2015.000000

           visit_day      visit_age     cmv_status    ebv_status  \
count  156118.000000  156118.000000  107540.000000  83087.000000
mean        0.063426      20.107988       0.360545      0.388280
std         0.317120      17.974787       0.480161      0.487362
min         0.000000       0.580000       0.000000      0.000000
25%         0.000000       7.280000       0.000000      0.000000
50%         0.000000      17.420000       0.000000      0.000000
75%         0.000000      25.030000       1.000000      1.000000
max         7.000000      90.000000       1.000000      1.000000

                bmi        vaccine  …  vaccinated_2yr_prior  \
count  61708.000000  103295.000000  …          53273.000000
mean      24.399225       3.419807  …              0.656806
std        5.432518       1.218462  …              0.879408
min       13.120000       1.000000  …              0.000000
25%       20.810000       3.000000  …              0.000000
50%       23.540000       4.000000  …              0.000000
75%       27.300000       4.000000  …              1.000000
```

```
max       52.120000       6.000000  …              3.000000

       vaccine_type_2yr_prior  vaccinated_3yr_prior  vaccine_type_3yr_prior  \
count            43979.000000          21706.000000            15855.000000
mean                 2.642284              0.901640                2.299401
std                  0.575715              0.856187                0.533698
min                  2.000000              0.000000                2.000000
25%                  2.000000              0.000000                2.000000
50%                  3.000000              1.000000                2.000000
75%                  3.000000              1.000000                3.000000
max                  4.000000              3.000000                4.000000

       vaccinated_4yr_prior  vaccine_type_4yr_prior  vaccinated_5yr_prior  \
count          21706.000000            15778.000000           8659.000000
mean               0.987653                2.348397              0.965008
std                0.946111                0.544284              0.951613
min                0.000000                2.000000              0.000000
25%                0.000000                2.000000              0.000000
50%                1.000000                2.000000              1.000000
75%                1.000000                3.000000              1.000000
max                3.000000                4.000000              3.000000

       vaccine_type_5yr_prior  influenza_infection_history  \
count             8254.000000                156118.000000
mean                 2.276957                     0.049860
std                  0.576707                     0.217656
min                  2.000000                     0.000000
25%                  2.000000                     0.000000
50%                  2.000000                     0.000000
75%                  2.000000                     0.000000
max                  4.000000                     1.000000

       influenza_hospitalization
count              156118.000000
mean                    0.004445
std                     0.066525
min                     0.000000
25%                     0.000000
50%                     0.000000
75%                     0.000000
max                     1.000000

[8 rows x 31 columns]
```

[3]:

```python
# Search for missing values
print(fludf.isnull().sum())
```

```
donor_id                            0
study_id                            0
gender                              0
race                              582
visit_id                            0
visit_year                          0
visit_day                           0
visit_type_hai                      0
visit_age                           0
cmv_status                      48578
ebv_status                      73031
bmi                             94410
vaccine                         52823
geo_mean                            0
d_geo_mean                      40806
vaccine_response                43662
mesurment_id                        0
assay                               0
name                                0
name_formatted                      0
subset                              0
units                               0
data                                0
statin_use                       2550
flu_vaccination_history        122588
total_vaccines_received        121888
vaccinated_1yr_prior             1182
vaccine_type_1yr_prior          24234
vaccinated_2yr_prior           102845
vaccine_type_2yr_prior         112139
vaccinated_3yr_prior           134412
vaccine_type_3yr_prior         140263
vaccinated_4yr_prior           134412
vaccine_type_4yr_prior         140340
vaccinated_5yr_prior           147459
vaccine_type_5yr_prior         147864
influenza_infection_history         0
influenza_hospitalization           0
dtype: int64
```

[4]:

```python
# Get column names
column_names = fludf.columns
print(column_names)
# Get column data types
print(fludf.dtypes)
```

```
Index(['donor_id', 'study_id', 'gender', 'race', 'visit_id', 'visit_year',
       'visit_day', 'visit_type_hai', 'visit_age', 'cmv_status', 'ebv_status',
       'bmi', 'vaccine', 'geo_mean', 'd_geo_mean', 'vaccine_response',
       'mesurment_id', 'assay', 'name', 'name_formatted', 'subset', 'units',
       'data', 'statin_use', 'flu_vaccination_history',
       'total_vaccines_received', 'vaccinated_1yr_prior',
       'vaccine_type_1yr_prior', 'vaccinated_2yr_prior',
       'vaccine_type_2yr_prior', 'vaccinated_3yr_prior',
       'vaccine_type_3yr_prior', 'vaccinated_4yr_prior',
       'vaccine_type_4yr_prior', 'vaccinated_5yr_prior',
       'vaccine_type_5yr_prior', 'influenza_infection_history',
       'influenza_hospitalization'],
      dtype='object')
donor_id                         int64
study_id                         int64
gender                          object
race                            object
visit_id                         int64
visit_year                       int64
visit_day                        int64
visit_type_hai                  object
visit_age                      float64
cmv_status                     float64
ebv_status                     float64
bmi                            float64
vaccine                        float64
geo_mean                       float64
d_geo_mean                     float64
vaccine_response               float64
mesurment_id                     int64
assay                            int64
name                            object
name_formatted                  object
subset                          object
units                           object
data                           float64
statin_use                     float64
flu_vaccination_history        float64
total_vaccines_received        float64
vaccinated_1yr_prior           float64
vaccine_type_1yr_prior         float64
vaccinated_2yr_prior           float64
vaccine_type_2yr_prior         float64
vaccinated_3yr_prior           float64
vaccine_type_3yr_prior         float64
vaccinated_4yr_prior           float64
vaccine_type_4yr_prior         float64
vaccinated_5yr_prior           float64
vaccine_type_5yr_prior         float64
influenza_infection_history      int64
```

```
influenza_hospitalization        int64
dtype: object
```

[18]:

```
fludf['name'].value_counts()
```

[18]:

```
L50_FASL                           555
L50_TNFB                           525
L50_MIP1A                          525
L50_MIG                            525
L50_IFNB                           525
L50_MIP1B                          525
L50_LIF                            525
L50_IFNG                           525
L50_IL12P40                        525
L50_LEPTIN                         525
L50_IL15                           525
L50_GCSF                           525
L50_TGFA                           525
L50_IL5                            525
L50_VEGF                           525
L50_IL10                           525
L50_FGFB                           525
L50_GMCSF                          525
L50_SCF                            525
L50_IL1B                           525
L50_IL7                            525
L50_IL1A                           525
L50_VCAM1                          525
L50_ICAM1                          525
L50_IL17                           525
L50_TNFA                           525
L50_MCP3                           525
L50_IL1RA                          525
L50_TGFB                           525
L50_IL17F                          525
                                  …
IFNa_EM CD8+ T cells: pSTAT3        20
LPS_EM CD4+ T cells: pErk1_2        20
LPS_EM CD4+ T cells: pSTAT5         20
IL-21_EM CD8+ T cells: pSTAT1       19
LPS_EM CD8+ T cells: pPLCg2         19
IL-21_EM CD8+ T cells: pSTAT5       19
LPS_EM CD8+ T cells: pErk1_2        19
IL-21_EM CD8+ T cells: IkBtot       19
IL-21_EM CD8+ T cells: pp38         19
LPS_EM CD8+ T cells: pp38           19
LPS_EM CD8+ T cells: pSTAT1         19
```

```
IL-21_EM CD8+ T cells: Ki67         19
IL-21_EM CD8+ T cells: pPLCg2       19
LPS_EM CD8+ T cells: pCREB          19
LPS_EM CD8+ T cells: Ki67           19
LPS_EM CD8+ T cells: IkBtot         19
IL-21_EM CD8+ T cells: pCREB        19
LPS_EM CD8+ T cells: pSTAT5         19
IL-21_EM CD8+ T cells: pErk1_2      19
LPS_EM CD8+ T cells: pSTAT3         19
IL-21_EM CD8+ T cells: pSTAT3       19
Unstim_EM CD8+ T cells: pp38        18
Unstim_EM CD8+ T cells: pPLCg2      18
Unstim_EM CD8+ T cells: pSTAT5      18
Unstim_EM CD8+ T cells: pCREB       18
Unstim_EM CD8+ T cells: IkBtot      18
Unstim_EM CD8+ T cells: pErk1_2     18
Unstim_EM CD8+ T cells: pSTAT1      18
Unstim_EM CD8+ T cells: pSTAT3      18
Unstim_EM CD8+ T cells: Ki67        18
Name: name, Length: 3283, dtype: int64
```

[16]:

```
fludf['subset'].value_counts()
```

[16]:

```
CD4+: pSTAT3                                              1288
CD8+: pSTAT1                                              1288
Mono: pSTAT5                                              1288
Mono: pSTAT1                                              1288
CD8+: pSTAT5                                              1288
B cell: pSTAT1                                            1288
CD4+: pSTAT1                                              1288
Mono: pSTAT3                                              1288
CD4+: pSTAT5                                              1288
B cell: pSTAT5                                            1288
CD8+: pSTAT3                                              1288
B cell: pSTAT3                                            1288
CD4+CD45RA+: pSTAT1                                       1050
CD8+CD45RA-: pSTAT3                                       1050
CD4+CD45RA-: pSTAT3                                       1050
CD8+CD45RA-: pSTAT5                                       1050
CD4+CD45RA+: pSTAT3                                       1050
CD4+CD45RA-: pSTAT1                                       1050
CD8+CD45RA+: pSTAT1                                       1050
CD8+CD45RA-: pSTAT1                                       1050
CD4+CD45RA+: pSTAT5                                       1050
CD4+CD45RA-: pSTAT5                                       1050
CD8+CD45RA+: pSTAT5                                       1050
CD8+CD45RA+: pSTAT3                                       1050
```

```
IL1B                                                       800
TNFA                                                       800
IL8                                                        800
IL6                                                        800
IFNG                                                       734
IL10                                                       734
                                                          …
NK-NKT: Lymph/CD3-/CD16+/CD56+/Q1: CD314-CD94+              26
B cell: Lymph/CD3-/CD19+CD20+/Q4: IgD-CD27-                 26
NK-NKT: Lymph/CD3+/CD8+/Q3: CD314+CD94-                     26
CXCR3 FMO: Lymph/CD3+/CD4+                                  26
NK-NKT: Lymph/CD3+/CD8-/Q3: CD314+CD94-                     26
PROGESTERONE                                                26
Treg: Lymph/CD3+/CD4+/CD25hiCD127low                        26
B cell: Lymph/CD3-/CD19+CD20+/CD24+/CD38-                   26
Treg: Lymph/CD3+/CD4+/CD25hiCD127low/Q1: CD161-CD45RA+      26
NK-NKT: Lymph/CD3+/CD8-                                     26
T cell: Lymph/CD3+/CD4+/Q2: CD45RA+CD27+                    26
T cell: Lymph/CD3+/CD4+/CD28+                               26
Treg: Lymph/CD3+/CD8+/CD161+                                26
CXCR3 FMO: Lymph/CD16+/CD56+                                26
NK-NKT: Lymph/CD3-/CD16+/CD56+/HLADR+                       26
Activated T: Lymph/CD3+/CD4+/Q1: HLADR-CD38+                26
CXCR3 FMO: Mono                                             26
B cell: Lymph/CD3-/CD19+CD20+                               26
Treg: Lymph/CD3+/CD8+                                       26
B cell: Lymph/CD3-/CD19+CD20+/Q3: IgD+CD27-                 26
Activated T: Lymph/CD3+                                     26
Activated T: Lymph/CD3+/CD4+                                26
NK-NKT: Lymph/CD3+CD56+                                     26
NK-NKT: Lymph/CD3-/CD16+/CD56+                              26
B cell: Lymph/CD3-                                          26
CXCR3 FMO: Mono/CD33+                                       26
CXCR3: Lymph/CD3+/CD4+/CXCR3+                               26
T cell: Lymph/CD3+/CD8+/Q4: CD45RA-CD27-                    26
B cell: Lymph/CD3-/CD20-                                    26
CXCR3: Lymph/CD3-                                           26
Name: subset, Length: 632, dtype: int64
```

[7]:

```
flusub = fludf[["donor_id", "gender","race", "visit_id", "visit_year",␣
 ↪"visit_day", "visit_age", "cmv_status", "ebv_status", "bmi", "statin_use",␣
 ↪"vaccine", "vaccine_response", "influenza_infection_history",␣
 ↪"influenza_hospitalization"]]
flusub.head()
```

[7]:

```
   donor_id  gender       race  visit_id  visit_year  visit_day  visit_age  \
0       813  Female  Caucasian      2937        2014          0       23.0
```

```
1       813  Female  Caucasian       2937        2014          0        23.0
2       813  Female  Caucasian       2937        2014          0        23.0
3       813  Female  Caucasian       2937        2014          0        23.0
4       813  Female  Caucasian       2937        2014          0        23.0

   cmv_status  ebv_status  bmi  statin_use  vaccine  vaccine_response  \
0         0.0         0.0  NaN         0.0      4.0               0.0
1         0.0         0.0  NaN         0.0      4.0               0.0
2         0.0         0.0  NaN         0.0      4.0               0.0
3         0.0         0.0  NaN         0.0      4.0               0.0
4         0.0         0.0  NaN         0.0      4.0               0.0

   influenza_infection_history  influenza_hospitalization
0                            0                          0
1                            0                          0
2                            0                          0
3                            0                          0
4                            0                          0
```

[8]:

```
# Search for missing values
print(flusub.isnull().sum())
```

```
donor_id                            0
gender                              0
race                              582
visit_id                            0
visit_year                          0
visit_day                           0
visit_age                           0
cmv_status                      48578
ebv_status                      73031
bmi                             94410
statin_use                       2550
vaccine                         52823
vaccine_response                43662
influenza_infection_history         0
influenza_hospitalization           0
dtype: int64
```

[9]:

```
result_flu = flusub.drop_duplicates()
result_flu.head()
```

[9]:

```
     donor_id  gender       race  visit_id  visit_year  visit_day  visit_age  \
0         813  Female  Caucasian      2937        2014          0       23.0
140       812    Male  Caucasian      2936        2014          0       28.0
280       811    Male  Caucasian      2935        2014          0       23.0
420       810    Male  Caucasian      2934        2014          0       27.0
```

```
560       809  Female       Asian      2933        2014          0       27.0

     cmv_status  ebv_status  bmi  statin_use  vaccine  vaccine_response  \
0           0.0         0.0  NaN         0.0      4.0               0.0
140         1.0         1.0  NaN         0.0      4.0               0.0
280         0.0         0.0  NaN         0.0      4.0               0.0
420         1.0         1.0  NaN         0.0      4.0               0.0
560         1.0         1.0  NaN         0.0      4.0               0.0

     influenza_infection_history  influenza_hospitalization
0                              0                          0
140                            0                          0
280                            0                          0
420                            0                          0
560                            0                          0
```

[10]:
```
result_flu.describe()
```

[10]:
```
        donor_id     visit_id   visit_year   visit_day   visit_age  \
count  740.00000   740.000000   740.000000  740.000000  740.000000
mean   426.57027  1026.764865  2010.787838    0.121622   38.354784
std    227.35277   796.029879     2.126478    0.526759   25.180895
min      1.00000     1.000000  2007.000000    0.000000    0.580000
25%    233.75000   372.250000  2009.000000    0.000000   19.775000
50%    421.50000   941.500000  2011.000000    0.000000   27.180000
75%    628.25000  1436.750000  2012.000000    0.000000   61.000000
max    813.00000  2937.000000  2015.000000    7.000000   90.000000

       cmv_status  ebv_status         bmi  statin_use     vaccine  \
count  311.000000  180.000000  475.000000  690.000000  619.000000
mean     0.411576    0.572222   24.984000    0.115942    3.626817
std      0.492912    0.496137    5.649661    0.320388    1.366967
min      0.000000    0.000000   13.120000    0.000000    1.000000
25%      0.000000    0.000000   21.245000    0.000000    4.000000
50%      0.000000    1.000000   24.050000    0.000000    4.000000
75%      1.000000    1.000000   28.395000    0.000000    4.000000
max      1.000000    1.000000   52.120000    1.000000    6.000000

       vaccine_response  influenza_infection_history  \
count        363.000000                   740.000000
mean           0.305785                     0.091892
std            0.461375                     0.289069
min            0.000000                     0.000000
25%            0.000000                     0.000000
50%            0.000000                     0.000000
75%            1.000000                     0.000000
max            1.000000                     1.000000
```

```
       influenza_hospitalization
count                 740.000000
mean                    0.010811
std                     0.103481
min                     0.000000
25%                     0.000000
50%                     0.000000
75%                     0.000000
max                     1.000000
```

[11]:
```
# Search for missing values
print(result_flu.isnull().sum())
```

```
donor_id                       0
gender                         0
race                           5
visit_id                       0
visit_year                     0
visit_day                      0
visit_age                      0
cmv_status                   429
ebv_status                   560
bmi                          265
statin_use                    50
vaccine                      121
vaccine_response             377
influenza_infection_history    0
influenza_hospitalization      0
dtype: int64
```

[12]:
```
result_flu.to_csv(r'C:/Users/Christine/Documents/Bellevue/DSC 680/Project 1/
 ↪result_flu.csv', header = True, index = False)
```

[ ]:
```
```