Data Cleaning with Python and Pandas Assignment
Intro to Data Science **Name: Robert Daniels**

Download RawDataForAssignment.csv from Canvas. Use Python and Pandas to clean the data and save to a new file.

Once you are finished, there should be

- no null values.
- no unrealistic ages or birthday months.
- only two categories for gender.

Pivot Check

Row Labels	Min of age	Max of age
Female	**1**	**84**
1	4	66
2	2	83
3	17	66
4	15	73
5	7	81
6	7	82
7	1	84
8	44	66
9	4	83
10	2	66
11	4	60
12	19	75
Male	**1**	**85**
1	7	78
2	1	84
3	1	79
4	18	70
5	1	85
6	3	60
7	3	73
8	12	80
9	5	85
10	3	78
11	2	78
12	9	83

1. Did you decide to delete rows or columns or replace the null values?

   Deleted email dimension, superfluous. Deleted null birthday_month dimension records, since mean does not make sense.
   Null accounted for ~11.4% of the age dimension, was unwilling to backfill that level of null, removed those values as well.

   What statements did you use to get rid of the null values?

   See screenshot at end

2. How did you handle the unrealistic ages and birthday months?

   Removed ages < 0 and > 122 (current world record)
   Removed birthday_month > 12 and < 1. Converted birthday_month to int for presentation

   What statements did you use?

   Screenshot

3. What statements did you use to consolidate the gender categories down to two?

   See screenshot

Type in your answers and save as a pdf file. Submit the pdf file along with your final CSV file.

dataclean.py

```python
'''Data Cleaning Assignment
Author: Robert D.
01/21/2022'''

import pandas as pd

# read in the data set
df = pd.read_csv("RawDataForAssignment.csv")

# get some initial impressions
print(df.dtypes)
print(df['gender'].value_counts())
print(df.describe())

# drop extraneous data
df = df.drop(['email'], axis=1)

# drop rows with null birthday_month (fill.mean makes no sense here)
df = df.dropna(subset=['birthday_month'])

'''Remove null age. Not willing to backfill ~11% of dataset with the mean.'''
df = df.dropna(subset=['age'])

# apply mask to filter nonsensical ages. 122 is world record
df = df[(df.age >= 0) & (df.age <= 122)]

# apply mask to birthday_month
df = df[(df.birthday_month >= 1) & (df.birthday_month <= 12)]

#convert column type
df = df.astype({'birthday_month' : 'int32'})

# clean gender. could also regex if contains ['f','F']
df['gender'] = df['gender'].str.strip()
df['gender'] = df['gender'].replace(['M','m','male'],'Male')
df['gender'] = df['gender'].replace(['F','F ''f','f ','female'],"Female")

# export clean data to csv
df.to_csv('scrubbed_data.csv')
```