Variable Encoding

February 6, 2021

Objective:

Perform preliminary steps for structuring the dataset:

1. The dataset contains 58 variables and over 800k observation points. Remove unwanted variable to reduce the size of the data set. The variables that can be immediately removed were identified in the previous week. Other can be removed along the way as necessary.
2. Provide data encoding for some key variables, such as the survey year and meal description.

[3]:
```
import pandas as pd

nhanes = pd.read_csv('../../Data/nhanes.csv')


#The following variables have been deemed irrelevant for this analysis, so they␣
 ↪are dropped.


nhanes = nhanes.drop(['DR1IVARA', 'DR1IVB12', 'DR1ICALC', 'DR1IIRON',␣
 ↪'DR1IZINC', 'DR1ISELE', 'DR1IP205',
             'DR1IP226', 'RIDRETH3', 'DR1I_PF_CUREDMEAT', 'DR1I_PF_ORGAN',␣
 ↪'DR1I_PF_POULT', 'DR1I_PF_MPS_TOTAL',
             'DR1I_PF_EGGS', 'DR1I_PF_NUTSDS', 'DR1I_PF_LEGUMES',␣
 ↪'DR1I_PF_TOTAL', 'DR1I_D_TOTAL',
             'DR1I_D_TOTAL', 'DR1I_D_MILK', 'DR1I_D_YOGURT', 'DR1I_D_CHEESE',␣
 ↪'WTDRD1_6YR'], axis=1)



#Map the survey year data, based on the SDDSRVYR encoding key

#Obtain description and value counts
nhanes['SDDSRVYR'].describe()
nhanes['SDDSRVYR'].value_counts()

#Create Survey Year variable based on lookup, mapping from CDC source
survey_year_lookup = {4: '2005-2006', 5: '2009-2010', 6: '2011-2012', 7:␣
 ↪'2013-2014', 8:'2015-2016', 9:'2017-2018'}
```

```
nhanes['Survey_Year'] = nhanes['SDDSRVYR'].map(survey_year_lookup)

#Check for NAs
print("Survey Year NA count is "+str(nhanes['Survey_Year'].isnull().sum()))


#Map the meal occasion data, based on the DR1.030Z encoding key

#Obtain description and value counts
nhanes['DR1.030Z'].describe()
nhanes['DR1.030Z'].value_counts()

#Create Survey Year variable based on lookup, mapping from CDC source
meal_name_lookup = {1: 'Breakfast', 2: 'Lunch', 3: 'Dinner', 4: 'Supper', 5:
 ↪'Brunch', 6:'Snack',
                    7: 'Drink', 8: 'Infant Feeding', 9: 'Extended consumption',␣
 ↪10: 'Desayano',
                    11: 'Almuerzo', 12: 'Comida', 13: 'Merienda', 14: 'Cena',␣
 ↪15: 'Enter comida',
                    16: 'Botana', 17: 'Bocadillo', 18: 'Tentempie', 19:␣
 ↪'Bebida', 91: 'Other'}


nhanes['Meal_Name'] = nhanes['DR1.030Z'].map(meal_name_lookup)

#Check for NAs
print("Meal Name NA count is "+str(nhanes['Meal_Name'].isnull().sum()))
```

```
Survey Year NA count is 0
Meal Name NA count is 0
```

[4]:
```
#Meal Name Counts - Observation Level
nhanes['Meal_Name'].value_counts()
```

[4]:
```
Dinner                  165082
Lunch                   161393
Breakfast               142660
Snack                   136295
Supper                   42739
Drink                    40487
Extended consumption     25242
Infant Feeding           18184
Cena                     18065
Desayano                 16198
Comida                   15428
Almuerzo                 13211
Merienda                  7026
```

```
Brunch                    6602
Bebida                    4958
Botana                    3291
Bocadillo                 2946
Enter comida              2842
Tentempie                  356
Other                        7
Name: Meal_Name, dtype: int64
```

[5]:
```
#Survey Name Counts - Observation Level
nhanes['Survey_Year'].value_counts()
```

[5]:
```
2011-2012    150991
2005-2006    146940
2009-2010    145703
2015-2016    131394
2013-2014    126503
2017-2018    121481
Name: Survey_Year, dtype: int64
```

[ ]:
```

```