pakistandhs2

May 8, 2024

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

[2]:
```
df= pd.read_stata('PKKR71FL.DTA')
```

1 Data Preprocessing

[3]:
```
df1=␣
 ↪df[['h11','caseid','v116','v113','b19','b4','m18','m19','v730','v012','v136','v106','v714',
```

[4]:
```
df1 = df1[df1['h11'] != "don't know"]
df1['h11'] = df1['h11'].map({'no': 0, 'yes, last two weeks': 1})
```

[5]:
```
df1= df1.dropna(subset=['h11'])
```

[6]:
```
df1
```

[6]:
```
       h11           caseid                                   v116  \
0      0.0         1   1  2                   flush to pit latrine
2      0.0         1   8  7                   flush to pit latrine
3      0.0         1   9  4                   flush to pit latrine
4      0.0         1   9  4                   flush to pit latrine
5      0.0         1  10  7                   flush to pit latrine
…    …              …                                    …
12703  0.0       580  21  2                   flush to pit latrine
12704  0.0       580  24  4                   flush to pit latrine
12705  0.0       580  25  3  ventilated improved pit latrine (vip)
12706  0.0       580  25  3  ventilated improved pit latrine (vip)
12707  0.0       580  28  3                   flush to septic tank

                                                    v113  b19      b4  \
0                                     unprotected spring   42    male
2      river/dam/lake/ponds/stream/canal/irrigation c…    8  female
3                                     unprotected spring   29  female
4                                     unprotected spring   49    male
5                                       protected spring   26  female
…                                                   …  …     …
```

```
12703                              tube well or borehole    9    male
12704                              tube well or borehole   17  female
12705                              tube well or borehole    1  female
12706                              tube well or borehole    1  female
12707                              tube well or borehole   40    male

                        m18                m19  v730  v012  v136  \
0      smaller than average  not weighed at birth  44.0    35     7
2                   average  not weighed at birth  25.0    21     9
3                   average  not weighed at birth  38.0    28     8
4                   average  not weighed at birth  38.0    28     8
5                   average                2000.0  45.0    35    11
…                     …                    …   …   …   …
12703   larger than average                5500.0  40.0    28     3
12704               average  not weighed at birth  30.0    25    11
12705               average                2400.0  33.0    25    10
12706               average                2200.0  33.0    25    10
12707               average            don't know  45.0    38     7

               v106 v714     v190 v101   v140
0         secondary   no  poorest  kpk  rural
2            higher   no   poorer  kpk  rural
3      no education   no  poorest  kpk  rural
4      no education   no  poorest  kpk  rural
5            higher  yes   poorer  kpk  rural
…             …  …      …  …    …
12703        higher  yes  richest  ajk  urban
12704     secondary   no   middle  ajk  urban
12705        higher   no  richest  ajk  urban
12706        higher   no  richest  ajk  urban
12707  no education   no   middle  ajk  urban

[11947 rows x 16 columns]
```

[7]:

```
df1.isnull().sum()
```

[7]:

```
h11         0
caseid      0
v116        0
v113        0
b19         0
b4          0
m18         3
m19         4
v730      131
v012        0
v136        0
```

```
v106        0
v714        2
v190        0
v101        0
v140        0
dtype: int64
```

[8]:
```
df1['v116'].value_counts()
```

[8]:
```
flush to septic tank                        3017
flush to pit latrine                        2868
flush to piped sewer system                 2659
no facility/bush/field                      1349
flush to somewhere else                      421
pit latrine with slab                        413
not a dejure resident                        380
pit latrine without slab/open pit            266
composting toilet                            238
bucket toilet                                107
flush, don't know where                      100
ventilated improved pit latrine (vip)         63
other                                         39
hanging toilet/latrine                        27
Name: v116, dtype: int64
```

[9]:
```
df1 = df1[df1['v116'] != "not a dejure resident"]
```

[10]:
```
# Define a list of improved toilet facility categories
improved_categories = ['flush to septic tank', 'flush to piped sewer system',
                       'flush to somewhere else', 'pit latrine with slab',␣
 ↪'composting toilet',
                       'flush, don\'t know where', 'ventilated improved pit␣
 ↪latrine (vip)']

# Create a binary variable indicating improved (1) and unimproved (0) toilet␣
 ↪facilities
df1['improved_toilet'] = df1['v116'].isin(improved_categories).astype(int)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_24536\3459918877.py:7:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df1['improved_toilet'] = df1['v116'].isin(improved_categories).astype(int)
```

[11]:
```
df1
```

```
[11]:        h11           caseid                                   v116  \
      0      0.0         1   1  2                   flush to pit latrine
      2      0.0         1   8  7                   flush to pit latrine
      3      0.0         1   9  4                   flush to pit latrine
      4      0.0         1   9  4                   flush to pit latrine
      5      0.0         1  10  7                   flush to pit latrine
      …    …              …                                    …
      12703  0.0       580  21  2                   flush to pit latrine
      12704  0.0       580  24  4                   flush to pit latrine
      12705  0.0       580  25  3  ventilated improved pit latrine (vip)
      12706  0.0       580  25  3  ventilated improved pit latrine (vip)
      12707  0.0       580  28  3                   flush to septic tank

                                                          v113  b19      b4  \
      0                                     unprotected spring   42    male
      2      river/dam/lake/ponds/stream/canal/irrigation c…    8  female
      3                                     unprotected spring   29  female
      4                                     unprotected spring   49    male
      5                                       protected spring   26  female
      …                                                  …  …     …
      12703                              tube well or borehole    9    male
      12704                              tube well or borehole   17  female
      12705                              tube well or borehole    1  female
      12706                              tube well or borehole    1  female
      12707                              tube well or borehole   40    male

                              m18                  m19  v730  v012  v136  \
      0      smaller than average  not weighed at birth  44.0    35     7
      2                   average  not weighed at birth  25.0    21     9
      3                   average  not weighed at birth  38.0    28     8
      4                   average  not weighed at birth  38.0    28     8
      5                   average                2000.0  45.0    35    11
      …                     …                   …   …   …   …
      12703   larger than average                5500.0  40.0    28     3
      12704               average  not weighed at birth  30.0    25    11
      12705               average                2400.0  33.0    25    10
      12706               average                2200.0  33.0    25    10
      12707               average            don't know  45.0    38     7

                     v106 v714     v190 v101   v140  improved_toilet
      0         secondary   no  poorest  kpk  rural                0
      2            higher   no   poorer  kpk  rural                0
      3      no education   no  poorest  kpk  rural                0
      4      no education   no  poorest  kpk  rural                0
      5            higher  yes   poorer  kpk  rural                0
      …             …  …      …  …    …              …
      12703        higher  yes  richest  ajk  urban                0
```

```
12704     secondary  no   middle  ajk  urban                0
12705        higher  no  richest  ajk  urban                1
12706        higher  no  richest  ajk  urban                1
12707  no education  no   middle  ajk  urban                1

[11567 rows x 17 columns]
```

[ ]:

[12]:

```
df1=df1.drop(columns=['v116'])
```

[13]:

```
# Define a list of improved drinking water source categories
improved_categories = ['piped into dwelling', 'public tap/standpipe',␣
 ↪'protected well', 'protected spring',
                       'piped to yard/plot', 'bottled water', 'filtration␣
 ↪plant', 'tube well or borehole']

# Create a binary variable indicating improved (1) and unimproved (0) drinking␣
 ↪water sources
df1['improved_water'] = df1['v113'].isin(improved_categories).astype(int)
df1=df1.drop(columns=['v113'])
```

[14]:

```
df1['improved_water'].value_counts()
```

[14]:

```
1    9925
0    1642
Name: improved_water, dtype: int64
```

[15]:

```
df1
```

[15]:

```
       h11           caseid  b19      b4                   m18  \
0      0.0         1   1  2   42    male  smaller than average
2      0.0         1   8  7    8  female               average
3      0.0         1   9  4   29  female               average
4      0.0         1   9  4   49    male               average
5      0.0         1  10  7   26  female               average
…    …              …  …     …                   …
12703  0.0       580  21  2    9    male   larger than average
12704  0.0       580  24  4   17  female               average
12705  0.0       580  25  3    1  female               average
12706  0.0       580  25  3    1  female               average
12707  0.0       580  28  3   40    male               average

                        m19  v730  v012  v136           v106 v714     v190  \
0      not weighed at birth  44.0    35     7      secondary   no  poorest
2      not weighed at birth  25.0    21     9         higher   no   poorer
3      not weighed at birth  38.0    28     8  no education   no  poorest
```

```
4      not weighed at birth  38.0    28     8  no education   no  poorest
5                    2000.0  45.0    35    11        higher  yes   poorer
…                        …    …     …    …            …   …       …
12703                5500.0  40.0    28     3        higher  yes  richest
12704  not weighed at birth  30.0    25    11     secondary   no   middle
12705                2400.0  33.0    25    10        higher   no  richest
12706                2200.0  33.0    25    10        higher   no  richest
12707            don't know  45.0    38     7  no education   no   middle

      v101   v140  improved_toilet  improved_water
0      kpk  rural                0               0
2      kpk  rural                0               0
3      kpk  rural                0               0
4      kpk  rural                0               0
5      kpk  rural                0               1
…     …    …               …              …
12703  ajk  urban                0               1
12704  ajk  urban                0               1
12705  ajk  urban                1               1
12706  ajk  urban                1               1
12707  ajk  urban                1               1

[11567 rows x 16 columns]
```

[16]:
```python
df1['b19'].info()
```

```
<class 'pandas.core.series.Series'>
Int64Index: 11567 entries, 0 to 12707
Series name: b19
Non-Null Count  Dtype
--------------  -----
11567 non-null  int8
dtypes: int8(1)
memory usage: 101.7 KB
```

[17]:
```python
df1['b4'].value_counts()
```

[17]:
```
male      5906
female    5661
Name: b4, dtype: int64
```

[18]:
```python
df1['Gender_male'] = df1['b4'].map({'female': 0, 'male': 1})
```

[19]:
```python
df1=df1.drop(columns=['b4'])
```

[20]:
```python
df1['m18'].value_counts()
```

[20]:
```
average                 8659
smaller than average    1515
larger than average      824
very small               458
very large                74
don't know                34
Name: m18, dtype: int64
```

[21]:
```
df1['Size_Child'] = df1['m18'].map({'very small': 1, 'smaller than average': 2,␣
↪'average': 3, 'larger than average': 4, 'very large': 5, "don't know": np.
↪nan })
```

[22]:
```
df1=df1.drop(columns=['m18'])
```

[23]:
```
df1
```

[23]:
```
       h11           caseid  b19                   m19  v730  v012  v136  \
0      0.0         1   1  2   42  not weighed at birth  44.0    35     7
2      0.0         1   8  7    8  not weighed at birth  25.0    21     9
3      0.0         1   9  4   29  not weighed at birth  38.0    28     8
4      0.0         1   9  4   49  not weighed at birth  38.0    28     8
5      0.0         1  10  7   26                2000.0  45.0    35    11
…    …              …  …                   …   …   …   …
12703  0.0       580  21  2    9                5500.0  40.0    28     3
12704  0.0       580  24  4   17  not weighed at birth  30.0    25    11
12705  0.0       580  25  3    1                2400.0  33.0    25    10
12706  0.0       580  25  3    1                2200.0  33.0    25    10
12707  0.0       580  28  3   40            don't know  45.0    38     7

               v106 v714     v190 v101    v140  improved_toilet  \
0          secondary   no  poorest  kpk  rural                0
2             higher   no   poorer  kpk  rural                0
3      no education   no  poorest  kpk  rural                0
4      no education   no  poorest  kpk  rural                0
5             higher  yes   poorer  kpk  rural                0
…             …  …      …  …    …              …
12703         higher  yes  richest  ajk  urban                0
12704      secondary   no   middle  ajk  urban                0
12705         higher   no  richest  ajk  urban                1
12706         higher   no  richest  ajk  urban                1
12707  no education   no   middle  ajk  urban                1

       improved_water Gender_male  Size_Child
0                   0           1         2.0
2                   0           0         3.0
3                   0           0         3.0
4                   0           1         3.0
```

```
5                   1           0         3.0
…               …         …         …
12703               1           1         4.0
12704               1           0         3.0
12705               1           0         3.0
12706               1           0         3.0
12707               1           1         3.0

[11567 rows x 16 columns]
```

[ ]:

```

```

[24]:

```
df1['m19'].replace({'not weighed at birth': np.nan, "don't know": np.nan},␣
 ↪inplace=True)
```

[25]:

```
df1['Birth Weight'] = df1['m19'].astype('float64')
```

[26]:

```
df1=df1.drop(columns=['m19'])
```

[27]:

```
df1['Birth Weight'].value_counts()
```

[27]:

```
3000.0    532
2500.0    303
3500.0    246
2000.0    185
4000.0    128
         …
3670.0      1
3540.0      1
750.0       1
3007.0      1
2980.0      1
Name: Birth Weight, Length: 111, dtype: int64
```

[28]:

```
df1.rename(columns={'v012': "mother's age"}, inplace=True)
```

[29]:

```
df1
```

[29]:

```
       h11           caseid  b19  v730  mother's age  v136          v106 v714  \
0      0.0         1   1  2   42  44.0            35     7     secondary   no
2      0.0         1   8  7    8  25.0            21     9        higher   no
3      0.0         1   9  4   29  38.0            28     8  no education   no
4      0.0         1   9  4   49  38.0            28     8  no education   no
5      0.0         1  10  7   26  45.0            35    11        higher  yes
…    …              …  …   …           …   …           …  …
12703  0.0       580  21  2    9  40.0            28     3        higher  yes
12704  0.0       580  24  4   17  30.0            25    11     secondary   no
12705  0.0       580  25  3    1  33.0            25    10        higher   no
```

```
12706  0.0        580  25  3    1  33.0              25    10        higher   no
12707  0.0        580  28  3   40  45.0              38     7  no education   no

          v190 v101   v140  improved_toilet  improved_water Gender_male  \
0      poorest  kpk  rural                0               0           1
2       poorer  kpk  rural                0               0           0
3      poorest  kpk  rural                0               0           0
4      poorest  kpk  rural                0               0           1
5       poorer  kpk  rural                0               1           0
…        …  …    …              …             …          …
12703  richest  ajk  urban                0               1           1
12704   middle  ajk  urban                0               1           0
12705  richest  ajk  urban                1               1           0
12706  richest  ajk  urban                1               1           0
12707   middle  ajk  urban                1               1           1

       Size_Child  Birth Weight
0             2.0           NaN
2             3.0           NaN
3             3.0           NaN
4             3.0           NaN
5             3.0        2000.0
…           …            …
12703         4.0        5500.0
12704         3.0           NaN
12705         3.0        2400.0
12706         3.0        2200.0
12707         3.0           NaN

[11567 rows x 16 columns]
```

[30]:
```
df1['v730']=df1['v730'].astype('float')
```

[31]:
```
df1['Age Difference Parents']= abs(df1['v730']-df1["mother's age"])
```

[32]:
```
df1
```

[32]:
```
       h11           caseid  b19  v730  mother's age  v136          v106 v714  \
0      0.0          1    1  2   42  44.0              35     7     secondary   no
2      0.0          1    8  7    8  25.0              21     9        higher   no
3      0.0          1    9  4   29  38.0              28     8  no education   no
4      0.0          1    9  4   49  38.0              28     8  no education   no
5      0.0          1   10  7   26  45.0              35    11        higher  yes
…    …              …  …   …           …   …           …  …
12703  0.0        580  21  2    9  40.0              28     3        higher  yes
12704  0.0        580  24  4   17  30.0              25    11     secondary   no
12705  0.0        580  25  3    1  33.0              25    10        higher   no
```

```
12706  0.0       580  25  3    1  33.0            25    10        higher   no
12707  0.0       580  28  3   40  45.0            38     7  no education   no

          v190 v101   v140  improved_toilet  improved_water Gender_male   \
0      poorest  kpk  rural                0               0           1
2       poorer  kpk  rural                0               0           0
3      poorest  kpk  rural                0               0           0
4      poorest  kpk  rural                0               0           1
5       poorer  kpk  rural                0               1           0
…        …  …    …              …            …         …
12703  richest  ajk  urban                0               1           1
12704   middle  ajk  urban                0               1           0
12705  richest  ajk  urban                1               1           0
12706  richest  ajk  urban                1               1           0
12707   middle  ajk  urban                1               1           1

       Size_Child  Birth Weight  Age Difference Parents
0             2.0           NaN                     9.0
2             3.0           NaN                     4.0
3             3.0           NaN                    10.0
4             3.0           NaN                    10.0
5             3.0        2000.0                    10.0
…           …           …                    …
12703         4.0        5500.0                    12.0
12704         3.0           NaN                     5.0
12705         3.0        2400.0                     8.0
12706         3.0        2200.0                     8.0
12707         3.0           NaN                     7.0

[11567 rows x 17 columns]
```

[33]:

```
df1=df1.drop(columns=['v730'])
```

[34]:

```
df1.rename(columns={'v136': "No. Household Members"}, inplace=True)
```

[35]:

```
df1
```

[35]:

```
       h11           caseid  b19  mother's age  No. Household Members   \
0      0.0         1   1  2   42            35                      7
2      0.0         1   8  7    8            21                      9
3      0.0         1   9  4   29            28                      8
4      0.0         1   9  4   49            28                      8
5      0.0         1  10  7   26            35                     11
…    …              …  …          …                    …
12703  0.0       580  21  2    9            28                      3
12704  0.0       580  24  4   17            25                     11
12705  0.0       580  25  3    1            25                     10
```

```
12706  0.0       580  25  3    1            25                     10
12707  0.0       580  28  3   40            38                      7

               v106 v714     v190 v101    v140  improved_toilet  \
0         secondary   no  poorest  kpk  rural                0
2            higher   no   poorer  kpk  rural                0
3      no education   no  poorest  kpk  rural                0
4      no education   no  poorest  kpk  rural                0
5            higher  yes   poorer  kpk  rural                0
…             …  …      …  …    …             …
12703        higher  yes  richest  ajk  urban                0
12704     secondary   no   middle  ajk  urban                0
12705        higher   no  richest  ajk  urban                1
12706        higher   no  richest  ajk  urban                1
12707  no education   no   middle  ajk  urban                1

       improved_water Gender_male  Size_Child  Birth Weight  \
0                   0           1         2.0           NaN
2                   0           0         3.0           NaN
3                   0           0         3.0           NaN
4                   0           1         3.0           NaN
5                   1           0         3.0        2000.0
…               …         …         …           …
12703               1           1         4.0        5500.0
12704               1           0         3.0           NaN
12705               1           0         3.0        2400.0
12706               1           0         3.0        2200.0
12707               1           1         3.0           NaN

       Age Difference Parents
0                         9.0
2                         4.0
3                        10.0
4                        10.0
5                        10.0
…                       …
12703                    12.0
12704                     5.0
12705                     8.0
12706                     8.0
12707                     7.0

[11567 rows x 16 columns]
```

[ ]:

```
```

[36]:

```
df1.rename(columns={'v714': "Mother's Working Status"}, inplace=True)
```

[37]:

```
df1["Mother's Working Status"].value_counts()
```

[37]:

```
no      10241
yes      1324
Name: Mother's Working Status, dtype: int64
```

[38]:

```
df1["Mother's Working Status"] = df1["Mother's Working Status"].map({'no': 0,␣
 ↪'yes': 1})
```

[39]:

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11567 entries, 0 to 12707
Data columns (total 16 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   h11                      11567 non-null  float64
 1   caseid                   11567 non-null  object
 2   b19                      11567 non-null  int8
 3   mother's age             11567 non-null  int8
 4   No. Household Members    11567 non-null  int8
 5   v106                     11567 non-null  category
 6   Mother's Working Status  11565 non-null  category
 7   v190                     11567 non-null  category
 8   v101                     11567 non-null  category
 9   v140                     11567 non-null  category
 10  improved_toilet          11567 non-null  int32
 11  improved_water           11567 non-null  int32
 12  Gender_male              11567 non-null  category
 13  Size_Child               11530 non-null  float64
 14  Birth Weight             2062 non-null   float64
 15  Age Difference Parents   11437 non-null  float64
dtypes: category(6), float64(4), int32(2), int8(3), object(1)
memory usage: 735.4+ KB
```

[40]:

```
df1
```

[40]:

```
       h11          caseid  b19  mother's age  No. Household Members  \
0      0.0         1    1  2   42            35                      7
2      0.0         1    8  7    8            21                      9
3      0.0         1    9  4   29            28                      8
4      0.0         1    9  4   49            28                      8
5      0.0         1   10  7   26            35                     11
…      …                …   …          …                    …
12703  0.0       580   21  2    9            28                      3
12704  0.0       580   24  4   17            25                     11
12705  0.0       580   25  3    1            25                     10
12706  0.0       580   25  3    1            25                     10
```

```
12707  0.0       580  28  3   40            38                      7

               v106 Mother's Working Status     v190 v101   v140  \
0         secondary                       0  poorest  kpk  rural
2            higher                       0   poorer  kpk  rural
3      no education                       0  poorest  kpk  rural
4      no education                       0  poorest  kpk  rural
5            higher                       1   poorer  kpk  rural
…             …                     …      …  …    …
12703        higher                       1  richest  ajk  urban
12704     secondary                       0   middle  ajk  urban
12705        higher                       0  richest  ajk  urban
12706        higher                       0  richest  ajk  urban
12707  no education                       0   middle  ajk  urban

       improved_toilet  improved_water Gender_male  Size_Child  Birth Weight  \
0                    0               0           1         2.0           NaN
2                    0               0           0         3.0           NaN
3                    0               0           0         3.0           NaN
4                    0               0           1         3.0           NaN
5                    0               1           0         3.0        2000.0
…                …             …         …         …           …
12703                0               1           1         4.0        5500.0
12704                0               1           0         3.0           NaN
12705                1               1           0         3.0        2400.0
12706                1               1           0         3.0        2200.0
12707                1               1           1         3.0           NaN

       Age Difference Parents
0                         9.0
2                         4.0
3                        10.0
4                        10.0
5                        10.0
…                       …
12703                    12.0
12704                     5.0
12705                     8.0
12706                     8.0
12707                     7.0

[11567 rows x 16 columns]
```

[ ]:

```
```

[41]:

```
df1
```

```
[41]:        h11           caseid  b19  mother's age  No. Household Members  \
      0      0.0         1   1  2   42            35                      7
      2      0.0         1   8  7    8            21                      9
      3      0.0         1   9  4   29            28                      8
      4      0.0         1   9  4   49            28                      8
      5      0.0         1  10  7   26            35                     11
      …    …              …  …           …                    …
      12703  0.0       580  21  2    9            28                      3
      12704  0.0       580  24  4   17            25                     11
      12705  0.0       580  25  3    1            25                     10
      12706  0.0       580  25  3    1            25                     10
      12707  0.0       580  28  3   40            38                      7

                     v106 Mother's Working Status     v190 v101   v140  \
      0         secondary                       0  poorest  kpk  rural
      2            higher                       0   poorer  kpk  rural
      3      no education                       0  poorest  kpk  rural
      4      no education                       0  poorest  kpk  rural
      5            higher                       1   poorer  kpk  rural
      …             …                     …      …  …    …
      12703        higher                       1  richest  ajk  urban
      12704     secondary                       0   middle  ajk  urban
      12705        higher                       0  richest  ajk  urban
      12706        higher                       0  richest  ajk  urban
      12707  no education                       0   middle  ajk  urban

             improved_toilet  improved_water Gender_male  Size_Child  Birth Weight  \
      0                    0               0           1         2.0           NaN
      2                    0               0           0         3.0           NaN
      3                    0               0           0         3.0           NaN
      4                    0               0           1         3.0           NaN
      5                    0               1           0         3.0        2000.0
      …                …             …         …         …           …
      12703                0               1           1         4.0        5500.0
      12704                0               1           0         3.0           NaN
      12705                1               1           0         3.0        2400.0
      12706                1               1           0         3.0        2200.0
      12707                1               1           1         3.0           NaN

             Age Difference Parents
      0                         9.0
      2                         4.0
      3                        10.0
      4                        10.0
      5                        10.0
      …                       …
      12703                    12.0
```

```
12704                     5.0
12705                     8.0
12706                     8.0
12707                     7.0

[11567 rows x 16 columns]
```

[42]:
```
df1['v190'].value_counts()
```

[42]:
```
poorest    2655
poorer     2633
middle     2282
richer     2025
richest    1972
Name: v190, dtype: int64
```

[43]:
```
df1['Wealth'] = df1['v190'].map({'poorest': 1, 'poorer': 2, 'middle':3,␣
 ↪'richer': 4, 'richest':5})
```

[44]:
```
df1=df1.drop(columns=['v190'])
```

[45]:
```
df1
```

[45]:
```
       h11          caseid  b19  mother's age  No. Household Members  \
0      0.0         1   1  2   42           35                      7
2      0.0         1   8  7    8           21                      9
3      0.0         1   9  4   29           28                      8
4      0.0         1   9  4   49           28                      8
5      0.0         1  10  7   26           35                     11
…    …              …  …            …                     …
12703  0.0       580  21  2    9           28                      3
12704  0.0       580  24  4   17           25                     11
12705  0.0       580  25  3    1           25                     10
12706  0.0       580  25  3    1           25                     10
12707  0.0       580  28  3   40           38                      7

               v106 Mother's Working Status v101   v140  improved_toilet  \
0         secondary                       0  kpk  rural                0
2            higher                       0  kpk  rural                0
3      no education                       0  kpk  rural                0
4      no education                       0  kpk  rural                0
5            higher                       1  kpk  rural                0
…             …                   …  …    …            …
12703        higher                       1  ajk  urban                0
12704     secondary                       0  ajk  urban                0
12705        higher                       0  ajk  urban                1
12706        higher                       0  ajk  urban                1
```

```
12707  no education                       0  ajk  urban                1

       improved_water Gender_male  Size_Child  Birth Weight  \
0                   0           1         2.0           NaN
2                   0           0         3.0           NaN
3                   0           0         3.0           NaN
4                   0           1         3.0           NaN
5                   1           0         3.0        2000.0
…               …         …         …           …
12703               1           1         4.0        5500.0
12704               1           0         3.0           NaN
12705               1           0         3.0        2400.0
12706               1           0         3.0        2200.0
12707               1           1         3.0           NaN

       Age Difference Parents Wealth
0                         9.0      1
2                         4.0      2
3                        10.0      1
4                        10.0      1
5                        10.0      2
…                       …    …
12703                    12.0      5
12704                     5.0      3
12705                     8.0      5
12706                     8.0      5
12707                     7.0      3

[11567 rows x 16 columns]
```

[46]:
```
df1['Residence_Urban'] = df1['v140'].map({'rural': 0, 'urban': 1})
df1['Residence_Urban'].replace('not a dejure resident', pd.NA, inplace=True)
```

[47]:
```
df1=df1.drop(columns=['v140'])
```

[48]:
```
df1.rename(columns={'v101': "Region"}, inplace=True)
```

[49]:
```
df1
```

[49]:
```
       h11           caseid  b19  mother's age  No. Household Members  \
0      0.0           1   1  2   42            35                     7
2      0.0           1   8  7    8            21                     9
3      0.0           1   9  4   29            28                     8
4      0.0           1   9  4   49            28                     8
5      0.0           1  10  7   26            35                    11
…    …              …  …           …                    …
12703  0.0         580  21  2    9            28                     3
```

```
12704  0.0       580  24  4   17            25                     11
12705  0.0       580  25  3    1            25                     10
12706  0.0       580  25  3    1            25                     10
12707  0.0       580  28  3   40            38                      7

               v106 Mother's Working Status Region  improved_toilet  \
0         secondary                       0    kpk                0
2            higher                       0    kpk                0
3      no education                       0    kpk                0
4      no education                       0    kpk                0
5            higher                       1    kpk                0
…             …                     …    …              …
12703        higher                       1    ajk                0
12704     secondary                       0    ajk                0
12705        higher                       0    ajk                1
12706        higher                       0    ajk                1
12707  no education                       0    ajk                1

       improved_water Gender_male  Size_Child  Birth Weight  \
0                   0           1         2.0           NaN
2                   0           0         3.0           NaN
3                   0           0         3.0           NaN
4                   0           1         3.0           NaN
5                   1           0         3.0        2000.0
…               …         …         …           …
12703               1           1         4.0        5500.0
12704               1           0         3.0           NaN
12705               1           0         3.0        2400.0
12706               1           0         3.0        2200.0
12707               1           1         3.0           NaN

       Age Difference Parents Wealth  Residence_Urban
0                         9.0      1              0.0
2                         4.0      2              0.0
3                        10.0      1              0.0
4                        10.0      1              0.0
5                        10.0      2              0.0
…                       …    …              …
12703                    12.0      5              1.0
12704                     5.0      3              1.0
12705                     8.0      5              1.0
12706                     8.0      5              1.0
12707                     7.0      3              1.0

[11567 rows x 16 columns]
```

[50]:

```
df1.rename(columns={'b19': "Age in Months"}, inplace=True)
df1.rename(columns={'v106': "Educational Attainment"}, inplace=True)
```

[51]:

```
dummy=pd.get_dummies(df1['Region'], prefix='Region',drop_first=True)
df1 = pd.concat([df1, dummy], axis=1)
df1=df1.drop(columns=['Region'])
```

[52]:

```
df1['Educational Attainment'] = df1['Educational Attainment'].map({'no␣
 ↪education': 1, 'primary': 2, 'secondary':3, 'higher': 4})
```

[53]:

```
df1.rename(columns={'h11': "Diarrhea Occurrence"}, inplace=True)
```

[54]:

```
#Filling the missing values
df1.isnull().sum()
```

[54]:

```
Diarrhea Occurrence          0
caseid                       0
Age in Months                0
mother's age                 0
No. Household Members        0
Educational Attainment       0
Mother's Working Status      2
improved_toilet              0
improved_water               0
Gender_male                  0
Size_Child                  37
Birth Weight              9505
Age Difference Parents     130
Wealth                       0
Residence_Urban              0
Region_sindh                 0
Region_kpk                   0
Region_balochistan           0
Region_gb                    0
Region_ict                   0
Region_ajk                   0
Region_fata                  0
dtype: int64
```

[55]:

```
median_age= df1['Age Difference Parents'].median()
```

[56]:

```
df1['Age Difference Parents'].fillna(median_age, inplace=True)
```

[57]:

```
df1
```

[57]:

```
       Diarrhea Occurrence          caseid  Age in Months  mother's age  \
0                      0.0         1    1  2             42            35
2                      0.0         1    8  7              8            21
```

```
3                      0.0         1   9  4             29            28
4                      0.0         1   9  4             49            28
5                      0.0         1  10  7             26            35
…                    …              …            …           …
12703                  0.0       580  21  2              9            28
12704                  0.0       580  24  4             17            25
12705                  0.0       580  25  3              1            25
12706                  0.0       580  25  3              1            25
12707                  0.0       580  28  3             40            38

       No. Household Members Educational Attainment Mother's Working Status  \
0                          7                      3                       0
2                          9                      4                       0
3                          8                      1                       0
4                          8                      1                       0
5                         11                      4                       1
…                      …                    …                     …
12703                      3                      4                       1
12704                     11                      3                       0
12705                     10                      4                       0
12706                     10                      4                       0
12707                      7                      1                       0

       improved_toilet  improved_water Gender_male  …  \
0                    0              0           1  …
2                    0              0           0  …
3                    0              0           0  …
4                    0              0           1  …
5                    0              1           0  …
…                …             …         …  …
12703                0              1           1  …
12704                0              1           0  …
12705                1              1           0  …
12706                1              1           0  …
12707                1              1           1  …

       Age Difference Parents  Wealth  Residence_Urban Region_sindh  \
0                         9.0       1              0.0            0
2                         4.0       2              0.0            0
3                        10.0       1              0.0            0
4                        10.0       1              0.0            0
5                        10.0       2              0.0            0
…                       …     …              …          …
12703                    12.0       5              1.0            0
12704                     5.0       3              1.0            0
12705                     8.0       5              1.0            0
12706                     8.0       5              1.0            0
```

```
12707                     7.0       3              1.0            0

       Region_kpk  Region_balochistan  Region_gb  Region_ict  Region_ajk  \
0               1                   0          0           0           0
2               1                   0          0           0           0
3               1                   0          0           0           0
4               1                   0          0           0           0
5               1                   0          0           0           0
…           …                 …        …         …         …
12703           0                   0          0           0           1
12704           0                   0          0           0           1
12705           0                   0          0           0           1
12706           0                   0          0           0           1
12707           0                   0          0           0           1

       Region_fata
0                0
2                0
3                0
4                0
5                0
…            …
12703            0
12704            0
12705            0
12706            0
12707            0

[11567 rows x 22 columns]
```

[58]:
```
mode_work= df1["Mother's Working Status"].mode()[0]
```

[59]:
```
df1["Mother's Working Status"].fillna(mode_work, inplace=True)
```

[60]:
```
median_size= df1['Size_Child'].median()
df1['Size_Child'].fillna(median_size, inplace=True)
```

[61]:
```
median_weight= df1['Birth Weight'].median()
df1['Birth Weight'].fillna(median_weight, inplace=True)
```

[62]:
```
df1['Educational Attainment'] = df1['Educational Attainment'].astype(float)
df1["Mother's Working Status"] = df1["Mother's Working Status"].astype(float)
df1["Gender_male"] = df1["Gender_male"].astype(float)
df1["Wealth"] = df1["Wealth"].astype(float)
```

[ ]:
```

```

2 Descriptive Statistics and EDA

[71]:
```
#Data Cleaning for Trend Analysis
df12= pd.read_stata('PKKR61FL.DTA')
```

[72]:
```
df12= df12[['h11','v116','v113','v101','v140']]
df12 = df12[df12['h11'] != "don't know"]
df12['h11'] = df12['h11'].map({'No': 0, 'Yes, last two weeks': 1})
df12=df12.dropna(subset=['h11'])
```

[73]:
```
# Define a list of improved drinking water source categories
improved_categories = ['Piped into dwelling', 'Public tap/standpipe',␣
 ↪'Protected well', 'Protected spring',
                       'Piped to yard/plot', 'Bottled water', 'Filtration␣
 ↪plant', 'Tube well or borehole']

# Create a binary variable indicating improved (1) and unimproved (0) drinking␣
 ↪water sources
df12['improved_water'] = df12['v113'].isin(improved_categories).astype(int)
df12=df12.drop(columns=['v113'])
df12.rename(columns={'h11': "Diarrhea Occurrence"}, inplace=True)
```

[74]:
```
# Define a list of improved toilet facility categories
improved_categories = ['Flush to septic tank', 'Flush to piped sewer system',
                       'Flush to somewhere else', 'Pit latrine with slab',
                       'Flush, don\'t know where', 'Ventilated Improved Pit␣
 ↪latrine (vip)']

# Create a binary variable indicating improved (1) and unimproved (0) toilet␣
 ↪facilities
df12['improved_toilet'] = df12['v116'].isin(improved_categories).astype(int)
df12=df12.drop(columns=['v116'])
```

[75]:
```
df12.rename(columns={'v140': "Place of Residence"}, inplace=True)
df12.rename(columns={'v101': "Region"}, inplace=True)
```

[76]:
```
df12['Region'].value_counts()
```

[76]:
```
Punjab                2994
Sindh                 2318
Khyber Pakhtunkhwa    2153
Balochistan           1730
Gilgit Baltistan      1013
Islamabad (ICT)        672
Name: Region, dtype: int64
```

[ ]:

[ ]:

```

```

[77]:

```
df2=df[['h11','v116','v113','v101','v140']]
df2 = df2[df2['h11'] != "don't know"]
df2['h11'] = df2['h11'].map({'no': 0, 'yes, last two weeks': 1})
df2=df2.dropna(subset=['h11'])
# Define a list of improved drinking water source categories
improved_categories = ['piped into dwelling', 'public tap/standpipe',␣
 ↪'protected well', 'protected spring',
                       'piped to yard/plot', 'bottled water', 'filtration␣
 ↪plant', 'tube well or borehole']

# Create a binary variable indicating improved (1) and unimproved (0) drinking␣
 ↪water sources
df2['improved_water'] = df2['v113'].isin(improved_categories).astype(int)
df2=df2.drop(columns=['v113'])
df2.rename(columns={'h11': "Diarrhea Occurrence"}, inplace=True)
```

[78]:

```
# Define a list of improved toilet facility categories
improved_categories = ['flush to septic tank', 'flush to piped sewer system',␣
 ↪'composting toilet'
                       'flush to somewhere else', 'pit latrine with slab',
                       'flush, don\'t know where', 'ventilated improved pit␣
 ↪latrine (vip)']

# Create a binary variable indicating improved (1) and unimproved (0) toilet␣
 ↪facilities
df2['improved_toilet'] = df2['v116'].isin(improved_categories).astype(int)
df2=df2.drop(columns=['v116'])
```

[79]:

```
df2.rename(columns={'v140': "Place of Residence"}, inplace=True)
df2.rename(columns={'v101': "Region"}, inplace=True)
```

[80]:

```
df2['Place of Residence'] = df2['Place of Residence'].map({'rural': 'Rural',␣
 ↪'urban': 'Urban'})
```

[81]:

```
df2
```

[81]:

```
       Diarrhea Occurrence Region Place of Residence  improved_water  \
0                      0.0    kpk              Rural               0
2                      0.0    kpk              Rural               0
3                      0.0    kpk              Rural               0
4                      0.0    kpk              Rural               0
5                      0.0    kpk              Rural               1
…                        …      …                  …               …
12703                  0.0    ajk              Urban               1
12704                  0.0    ajk              Urban               1
```

```
12705                  0.0    ajk              Urban                  1
12706                  0.0    ajk              Urban                  1
12707                  0.0    ajk              Urban                  1

       improved_toilet
0                    0
2                    0
3                    0
4                    0
5                    0
…                  …
12703                0
12704                0
12705                1
12706                1
12707                1

[11947 rows x 5 columns]
```

```
[82]: df2['Region'].value_counts()
```

```
[82]: punjab         2553
      sindh          2147
      kpk            1987
      balochistan    1379
      ajk            1247
      fata            999
      gb              864
      ict             771
      Name: Region, dtype: int64
```

```
[83]: df2['Region'] = df2['Region'].map({'punjab': 'Punjab', 'sindh': 'Sindh', 'kpk':␣
        ↪'Khyber Pakhtunkhwa', 'balochistan': 'Balochistan', 'ajk': 'Azad Jammu␣
        ↪Kashmir', 'gb':'Gilgit Baltistan', 'ict': 'Islamabad (ICT)', 'fata': 'Fata'})
```

```
[84]: df2= df2[(df2['Region']!='Azad Jammu Kashmir') & (df2['Region']!='Fata')]
```

```
[85]: df12
```

```
[85]:        Diarrhea Occurrence           Region Place of Residence  \
      0                      0.0           Punjab              Urban
      1                      0.0           Punjab              Urban
      2                      0.0           Punjab              Urban
      3                      0.0           Punjab              Urban
      4                      0.0           Punjab              Urban
      …                      …              …                 …
      11758                  0.0  Islamabad (ICT)              Rural
```

```
11759                  0.0  Islamabad (ICT)             Rural
11760                  0.0  Islamabad (ICT)             Rural
11761                  0.0  Islamabad (ICT)             Rural
11762                  0.0  Islamabad (ICT)             Rural

       improved_water  improved_toilet
0                   1                1
1                   1                1
2                   1                1
3                   1                1
4                   0                1
…                 …              …
11758               1                1
11759               1                1
11760               1                1
11761               1                1
11762               1                1

[10880 rows x 5 columns]
```

[86]:

```
# Add 'year' column to df2 and df12
df2['year'] = '2017-18'
df12['year'] = '2012-13'

# Concatenate df2 and df12
cdf = pd.concat([df2, df12], ignore_index=True)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_24536\234504397.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df2['year'] = '2017-18'
```

[ ]:

[87]:

```
cdf=cdf[cdf['Place of Residence']!= 'Not a dejure resident']
```

[ ]:

[88]:

```
#Descriptive Statistics
df1.describe().transpose()
```

[88]:

```
                           count          mean          std   min     25%  \
Diarrhea Occurrence      11567.0      0.175672     0.380557   0.0     0.0
Age in Months            11567.0     29.425175    17.502652   0.0    14.0
```

```
mother's age             11567.0    29.352295    6.084052   15.0    25.0
No. Household Members    11567.0     9.394744    5.009360    2.0     6.0
Educational Attainment   11567.0     1.973891    1.137454    1.0     1.0
Mother's Working Status  11567.0     0.114464    0.318387    0.0     0.0
improved_toilet          11567.0     0.597476    0.490428    0.0     0.0
improved_water           11567.0     0.858044    0.349020    0.0     1.0
Gender_male              11567.0     0.510590    0.499909    0.0     0.0
Size_Child               11567.0     2.873865    0.608528    1.0     3.0
Birth Weight             11567.0  2997.125184  350.543711  500.0  3000.0
Age Difference Parents   11567.0     5.388000    4.703583    0.0     2.0
Wealth                   11567.0     2.829342    1.404925    1.0     2.0
Residence_Urban          11567.0     0.440477    0.496466    0.0     0.0
Region_sindh             11567.0     0.180168    0.384344    0.0     0.0
Region_kpk               11567.0     0.168410    0.374246    0.0     0.0
Region_balochistan       11567.0     0.117489    0.322016    0.0     0.0
Region_gb                11567.0     0.071756    0.258094    0.0     0.0
Region_ict               11567.0     0.064407    0.245488    0.0     0.0
Region_ajk               11567.0     0.101582    0.302111    0.0     0.0
Region_fata              11567.0     0.086366    0.280916    0.0     0.0

                            50%     75%     max
Diarrhea Occurrence         0.0     0.0     1.0
Age in Months              29.0    45.0    59.0
mother's age               29.0    33.0    49.0
No. Household Members       8.0    11.0    44.0
Educational Attainment      1.0     3.0     4.0
Mother's Working Status     0.0     0.0     1.0
improved_toilet             1.0     1.0     1.0
improved_water              1.0     1.0     1.0
Gender_male                 1.0     1.0     1.0
Size_Child                  3.0     3.0     5.0
Birth Weight             3000.0  3000.0  6000.0
Age Difference Parents      4.0     7.0    51.0
Wealth                      3.0     4.0     5.0
Residence_Urban             0.0     1.0     1.0
Region_sindh                0.0     0.0     1.0
Region_kpk                  0.0     0.0     1.0
Region_balochistan          0.0     0.0     1.0
Region_gb                   0.0     0.0     1.0
Region_ict                  0.0     0.0     1.0
Region_ajk                  0.0     0.0     1.0
Region_fata                 0.0     0.0     1.0
```

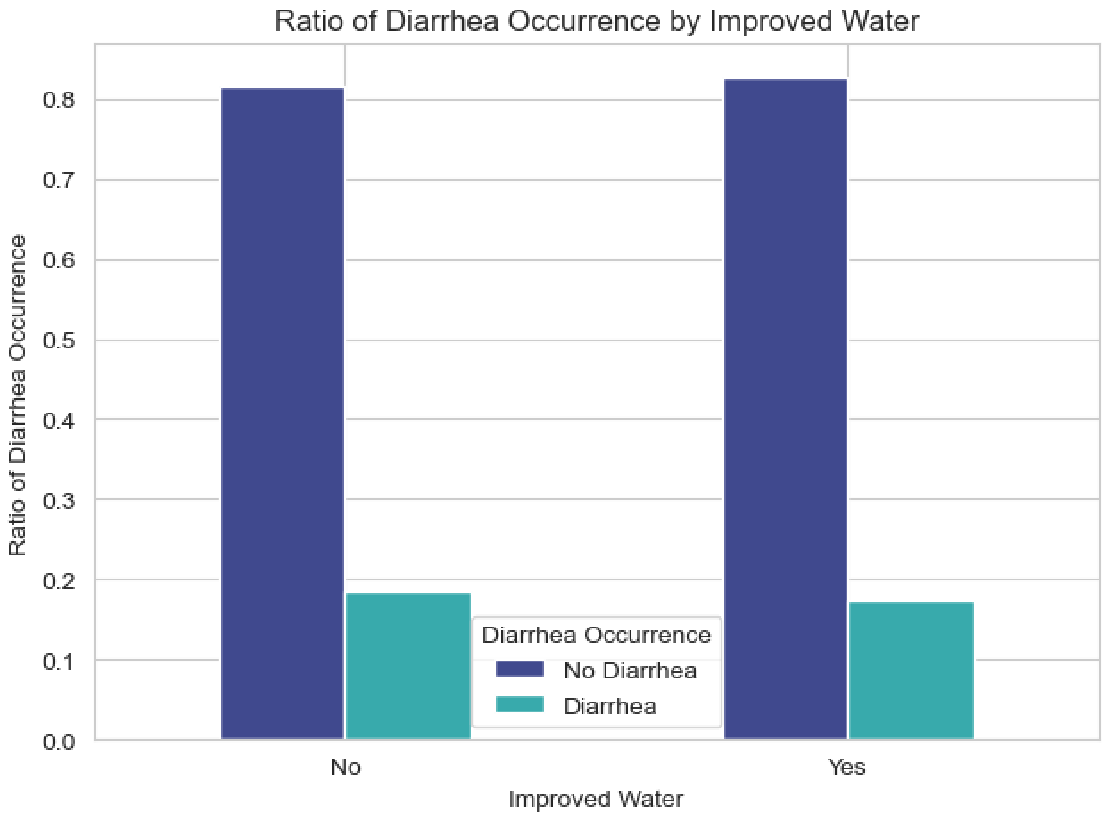
[118]:

```python
#EDA
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
# Calculate counts
counts = df1.groupby(['improved_water', 'Diarrhea Occurrence']).size().unstack()

# Calculate ratios
ratios = counts.div(counts.sum(axis=1), axis=0)
colors = sns.color_palette('mako', n_colors=ratios.shape[1])
# Plot countplot
ax = ratios.plot(kind='bar', color=colors)
plt.xlabel('Improved Water')
ax.set_xticklabels(['No', 'Yes'])

plt.ylabel('Ratio of Diarrhea Occurrence')
plt.title('Ratio of Diarrhea Occurrence by Improved Water')
plt.savefig("diarrheabywater.png", dpi=300)
plt.xticks(rotation=0)
plt.legend(title='Diarrhea Occurrence', labels=['No Diarrhea', 'Diarrhea'])
plt.tight_layout()
plt.show()
```



[94]:

```python
ratios
```

```
[94]: Diarrhea Occurrence       0.0       1.0
      improved_water
      0                    0.814860  0.185140
      1                    0.825894  0.174106
```

```
[95]: df2
```

```
[95]:        Diarrhea Occurrence              Region Place of Residence  \
      0                      0.0  Khyber Pakhtunkhwa              Rural
      2                      0.0  Khyber Pakhtunkhwa              Rural
      3                      0.0  Khyber Pakhtunkhwa              Rural
      4                      0.0  Khyber Pakhtunkhwa              Rural
      5                      0.0  Khyber Pakhtunkhwa              Rural
      …                    …                 …                …
      11383                  0.0    Gilgit Baltistan              Rural
      11384                  0.0    Gilgit Baltistan              Rural
      11385                  0.0    Gilgit Baltistan              Rural
      11386                  0.0    Gilgit Baltistan              Rural
      11387                  0.0    Gilgit Baltistan              Rural

             improved_water  improved_toilet     year
      0                   0                0  2017-18
      2                   0                0  2017-18
      3                   0                0  2017-18
      4                   0                0  2017-18
      5                   1                0  2017-18
      …                …              …      …
      11383               1                0  2017-18
      11384               1                1  2017-18
      11385               1                1  2017-18
      11386               1                0  2017-18
      11387               1                0  2017-18

      [9701 rows x 6 columns]
```

```
[96]: cdf
```

```
[96]:        Diarrhea Occurrence              Region Place of Residence  \
      0                      0.0  Khyber Pakhtunkhwa              Rural
      1                      0.0  Khyber Pakhtunkhwa              Rural
      2                      0.0  Khyber Pakhtunkhwa              Rural
      3                      0.0  Khyber Pakhtunkhwa              Rural
      4                      0.0  Khyber Pakhtunkhwa              Rural
      …                    …                 …                …
      20576                  0.0     Islamabad (ICT)              Rural
      20577                  0.0     Islamabad (ICT)              Rural
      20578                  0.0     Islamabad (ICT)              Rural
```

```
20579                  0.0     Islamabad (ICT)              Rural
20580                  0.0     Islamabad (ICT)              Rural

       improved_water  improved_toilet     year
0                   0                0  2017-18
1                   0                0  2017-18
2                   0                0  2017-18
3                   0                0  2017-18
4                   1                0  2017-18
…                …              …       …
20576               1                1  2012-13
20577               1                1  2012-13
20578               1                1  2012-13
20579               1                1  2012-13
20580               1                1  2012-13

[20184 rows x 6 columns]
```

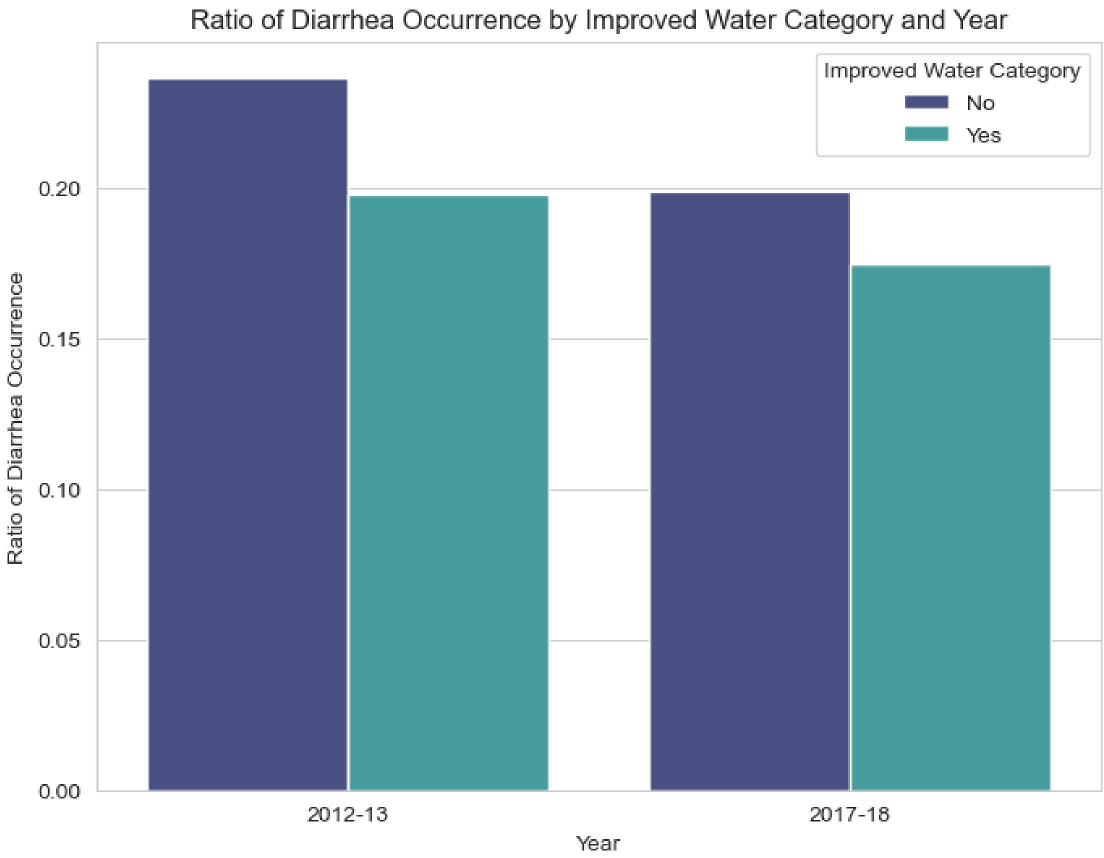
```
[ ]:
```

```
[109]: cdf['improved_water']=cdf['improved_water'].replace({0: 'No', 1: 'Yes'})
# Calculate the ratio of Diarrhea Occurrence equal to 1 for each year and␣
 ↪improved_water category
ratios = cdf.groupby(['year', 'improved_water'])['Diarrhea Occurrence'].mean().
 ↪reset_index()

# Set the style of the plot
sns.set_style("whitegrid")

# Create the bar plot
plt.figure(figsize=(8, 6))  # Set the figure size
sns.barplot(data=ratios, x='year', y='Diarrhea Occurrence',␣
 ↪hue='improved_water', palette='mako')

# Set the labels and title
plt.xlabel('Year')
plt.ylabel('Ratio of Diarrhea Occurrence')
plt.title('Ratio of Diarrhea Occurrence by Improved Water Category and Year')
plt.savefig("diarrheabywater.png", dpi=300)
# Show the plot
plt.legend(title='Improved Water Category')
plt.show()
```



```
[98]: ratios
```

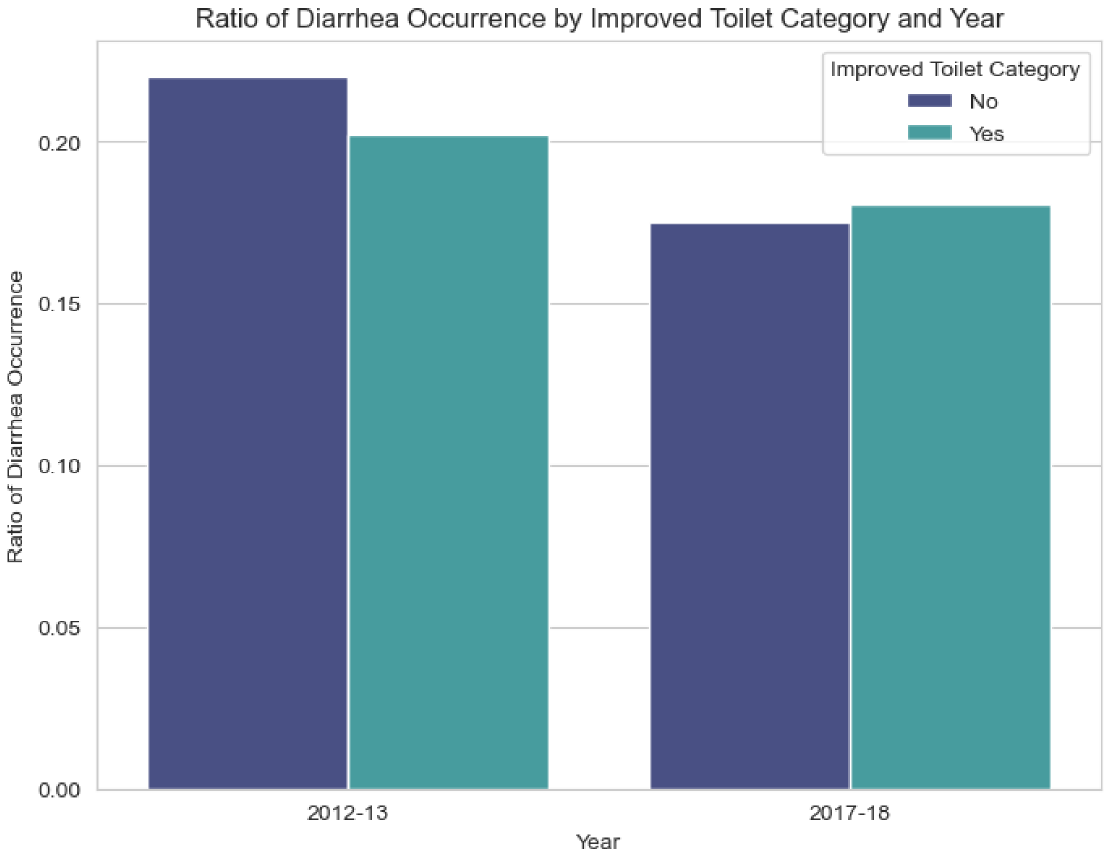
```
[98]:       year improved_water  Diarrhea Occurrence
      0  2012-13             No             0.236865
      1  2012-13            Yes             0.198210
      2  2017-18             No             0.199095
      3  2017-18            Yes             0.174806
```

```
[110]: import seaborn as sns
       import matplotlib.pyplot as plt
       cdf['improved_toilet']=cdf['improved_toilet'].replace({0: 'No', 1: 'Yes'})
       # Calculate the ratio of Diarrhea Occurrence equal to 1 for each year and␣
        ↪improved_toilet category
       ratios = cdf.groupby(['year', 'improved_toilet'])['Diarrhea Occurrence'].mean().
        ↪reset_index()

       # Set the style of the plot
       sns.set_style("whitegrid")
```

```
# Create the bar plot
plt.figure(figsize=(8, 6))  # Set the figure size
sns.barplot(data=ratios, x='year', y='Diarrhea Occurrence',␣
 ↪hue='improved_toilet', palette='mako')

# Set the labels and title
plt.xlabel('Year')
plt.ylabel('Ratio of Diarrhea Occurrence')
plt.title('Ratio of Diarrhea Occurrence by Improved Toilet Category and Year')
plt.savefig("diarrheabytoilet.png", dpi=300)
# Show the plot
plt.legend(title='Improved Toilet Category')
plt.show()
```



[100]:
```
ratios
```

[100]:
```
      year improved_toilet  Diarrhea Occurrence
0  2012-13              No             0.220418
1  2012-13             Yes             0.202484
```

```
2  2017-18                No             0.174964
3  2017-18               Yes             0.180478
```

[119]:

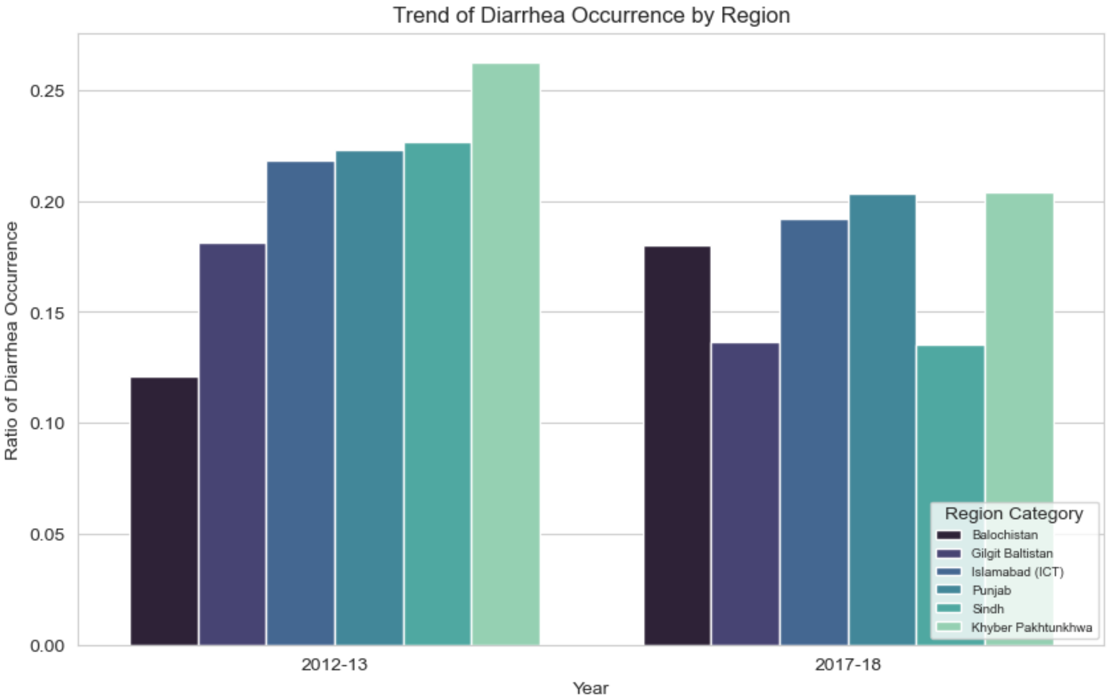
```
ratios = cdf.groupby(['year', 'Region'])['Diarrhea Occurrence'].mean().
 ↪reset_index()

# Sort the DataFrame by 'year' and 'Diarrhea Occurrence' in ascending order
ratios.sort_values(by=['year', 'Diarrhea Occurrence'], ascending=[True, True],␣
 ↪inplace=True)

# Set the style of the plot
sns.set_style("whitegrid")

# Create the bar plot
plt.figure(figsize=(10, 6))  # Set the figure size
sns.barplot(data=ratios, x='year', y='Diarrhea Occurrence', hue='Region',␣
 ↪palette='mako')

# Set the labels and title
plt.xlabel('Year')
plt.ylabel('Ratio of Diarrhea Occurrence')
plt.title('Trend of Diarrhea Occurrence by Region')
plt.savefig("diarrheabyregion.png", dpi=300)
# Show the plot
plt.legend(title='Region Category', loc='lower right', fontsize=7)
plt.show()
```



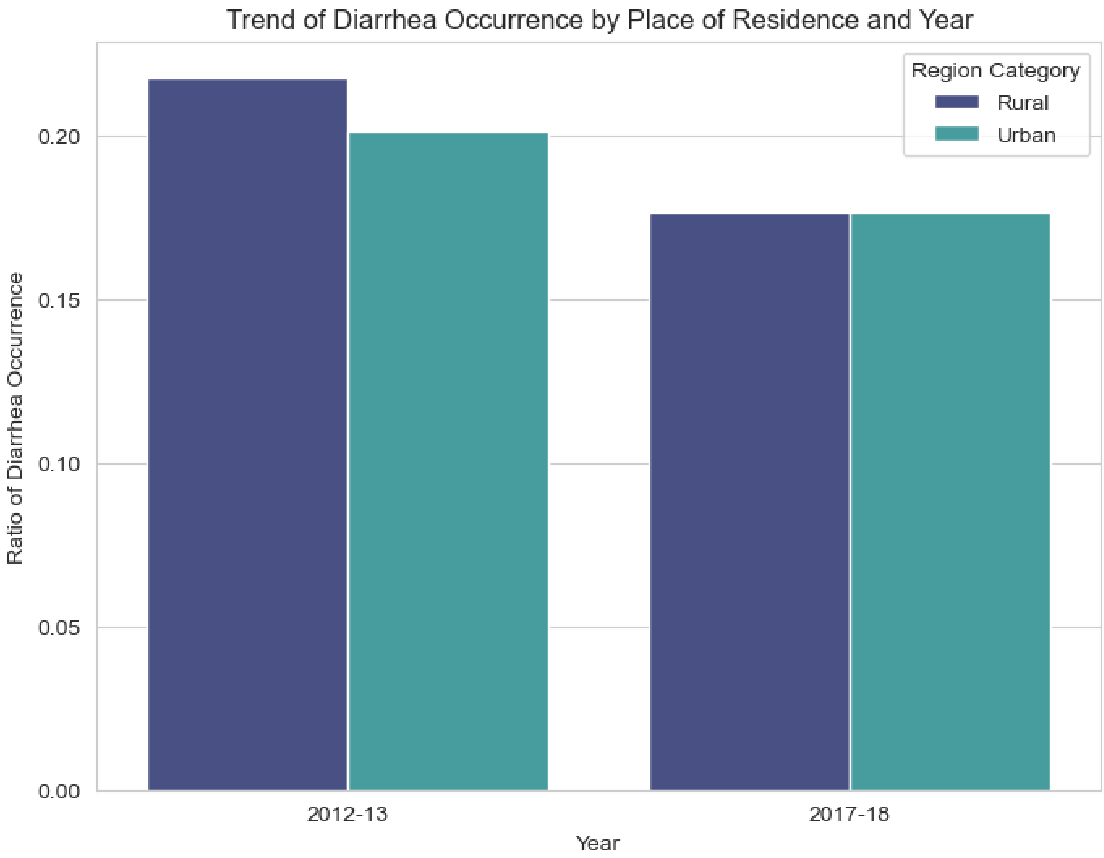
[112]:

```
# Calculate the ratio of Diarrhea Occurrence equal to 1 for each year and␣
 ↪improved_water category
ratios = cdf.groupby(['year', 'Place of Residence'])['Diarrhea Occurrence'].
 ↪mean().reset_index()

# Set the style of the plot
sns.set_style("whitegrid")

# Create the bar plot
plt.figure(figsize=(8, 6))  # Set the figure size
sns.barplot(data=ratios, x='year', y='Diarrhea Occurrence', hue='Place of␣
 ↪Residence', palette='mako')

# Set the labels and title
plt.xlabel('Year')
plt.ylabel('Ratio of Diarrhea Occurrence')
plt.title('Trend of Diarrhea Occurrence by Place of Residence and Year')
plt.savefig("diarrheabyresidence.png", dpi=300)
# Show the plot
plt.legend(title='Region Category', loc= 'upper right', fontsize=10)
plt.show()
```



[113]:

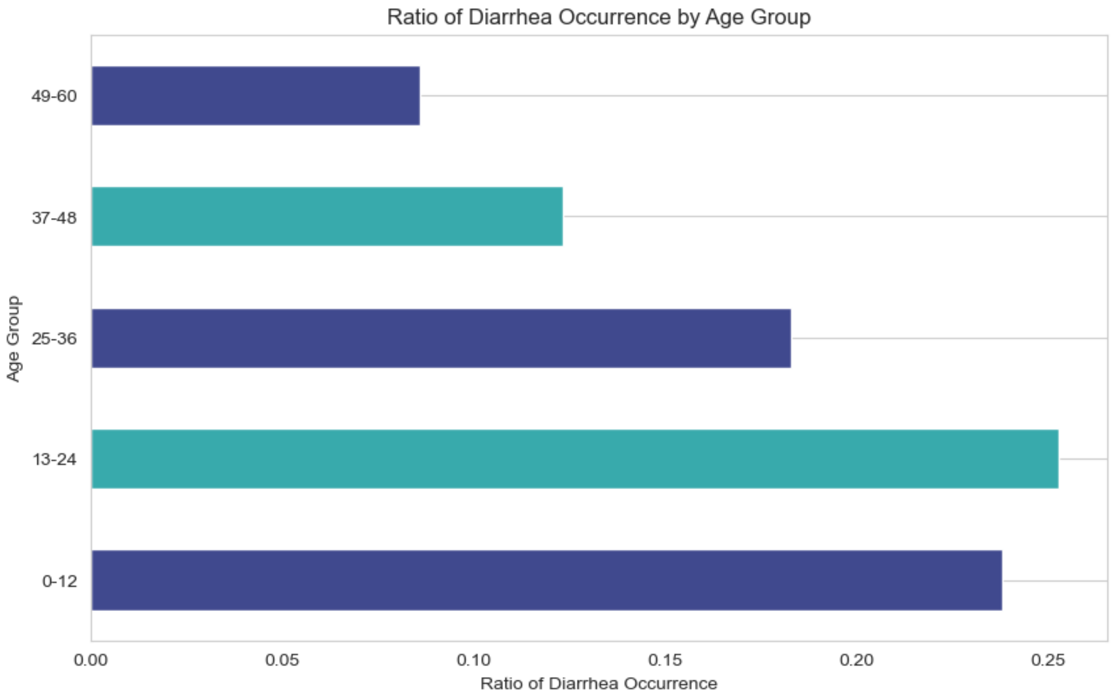
```
# Plot the line chart
# Define age bands
age_bins = [0, 12, 24, 36, 48, 60]
age_labels = ['0-12', '13-24', '25-36', '37-48', '49-60']

# Cut the Age in Months into age bands
df1['Age Group'] = pd.cut(df1['Age in Months'], bins=age_bins,␣
 ↪labels=age_labels, right=False)

# Calculate the mean Diarrhea Occurrence for each age group
age_diarrhea_mean = df1.groupby('Age Group')['Diarrhea Occurrence'].mean()

# Plot the horizontal bar chart
plt.figure(figsize=(10, 6))  # Set the figure size
age_diarrhea_mean.plot(kind='barh', color=colors)

# Add labels and title
plt.xlabel('Ratio of Diarrhea Occurrence')
plt.ylabel('Age Group')
plt.title('Ratio of Diarrhea Occurrence by Age Group')
plt.savefig("diarrheabyage.png", dpi=300)
# Display the plot
plt.grid(axis='x')  # Add gridlines only on x-axis
plt.show()
```



3 Cross Tabulation

[180]:
```
cross_tab = pd.crosstab(index=[df1['Diarrhea Occurrence']],␣
 ↪columns=df1['improved_toilet'])
```

[181]:
```
cross_tab
```

[181]:
```
improved_toilet         0     1
Diarrhea Occurrence
0.0                  3859  5676
1.0                   797  1235
```

[182]:
```
cross_tab1 = pd.crosstab(index=[df1['Diarrhea Occurrence']],␣
 ↪columns=df1['improved_water'])
```

[183]:
```
cross_tab1
```

[183]:
```
improved_water          0     1
Diarrhea Occurrence
0.0                  1338  8197
1.0                   304  1728
```

4 Running a Logistic Regression Model

[114]:
```
#Running a logistic regression model
import statsmodels.api as sm

# Define independent variables (numerical data)
X = df1.select_dtypes(include='number').drop(columns=['Diarrhea Occurrence'],␣
 ↪axis=1)

# Add constant term
X = sm.add_constant(X)

# Define dependent variable
y = df1['Diarrhea Occurrence']

# Fit logistic regression model
logit_model = sm.Logit(y, X, max_iter=50)
result = logit_model.fit()

# Display summary statistics and regression coefficients
print(result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.447779
         Iterations 6
                            Logit Regression Results
===============================================================================
Dep. Variable:     Diarrhea Occurrence   No. Observations:                11567
Model:                           Logit   Df Residuals:                    11546
Method:                            MLE   Df Model:                           20
Date:                 Wed, 08 May 2024   Pseudo R-squ.:                 0.03655
Time:                         23:11:50   Log-Likelihood:                -5179.5
converged:                        True   LL-Null:                       -5376.0
Covariance Type:             nonrobust   LLR p-value:                 5.784e-71
================================================================================
===========
                              coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-----------
const                      -0.5873      0.276     -2.130      0.033      -1.128
-0.047
```

```
Age in Months              -0.0236      0.002    -15.735      0.000      -0.027
-0.021
mother's age               -0.0102      0.004     -2.393      0.017      -0.018
-0.002
No. Household Members       0.0021      0.005      0.408      0.684      -0.008
0.012
Educational Attainment      0.0560      0.029      1.910      0.056      -0.001
0.114
Mother's Working Status    -0.0569      0.083     -0.683      0.495      -0.220
0.106
improved_toilet             0.0956      0.061      1.579      0.114      -0.023
0.214
improved_water             -0.0557      0.075     -0.747      0.455      -0.202
0.090
Gender_male                 0.1028      0.050      2.063      0.039       0.005
0.200
Size_Child                 -0.0671      0.042     -1.615      0.106      -0.148
0.014
Birth Weight                0.0001   7.21e-05      1.691      0.091   -1.94e-05
0.000
Age Difference Parents      0.0107      0.005      2.079      0.038       0.001
0.021
Wealth                     -0.1024      0.029     -3.560      0.000      -0.159
-0.046
Residence_Urban             0.0827      0.060      1.386      0.166      -0.034
0.200
Region_sindh               -0.5311      0.086     -6.190      0.000      -0.699
-0.363
Region_kpk                 -0.0086      0.080     -0.107      0.915      -0.166
0.149
Region_balochistan         -0.1650      0.095     -1.734      0.083      -0.352
0.022
Region_gb                  -0.5743      0.119     -4.808      0.000      -0.809
-0.340
Region_ict                 -0.0569      0.110     -0.515      0.606      -0.273
0.160
Region_ajk                 -0.3876      0.102     -3.806      0.000      -0.587
-0.188
Region_fata                -0.0712      0.104     -0.684      0.494      -0.275
0.133
================================================================================
===========

D:\Anaconda3\Lib\site-packages\statsmodels\base\model.py:130: ValueWarning:
unknown kwargs ['max_iter']
  warnings.warn(msg, ValueWarning)
D:\Anaconda3\Lib\site-packages\statsmodels\base\model.py:130: ValueWarning:
unknown kwargs ['max_iter']
```

```
  warnings.warn(msg, ValueWarning)
```

[65]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11567 entries, 0 to 12707
Data columns (total 22 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Diarrhea Occurrence     11567 non-null  float64
 1   caseid                  11567 non-null  object
 2   Age in Months           11567 non-null  int8
 3   mother's age            11567 non-null  int8
 4   No. Household Members   11567 non-null  int8
 5   Educational Attainment  11567 non-null  float64
 6   Mother's Working Status 11567 non-null  float64
 7   improved_toilet         11567 non-null  int32
 8   improved_water          11567 non-null  int32
 9   Gender_male             11567 non-null  float64
 10  Size_Child              11567 non-null  float64
 11  Birth Weight            11567 non-null  float64
 12  Age Difference Parents  11567 non-null  float64
 13  Wealth                  11567 non-null  float64
 14  Residence_Urban         11567 non-null  float64
 15  Region_sindh            11567 non-null  uint8
 16  Region_kpk              11567 non-null  uint8
 17  Region_balochistan      11567 non-null  uint8
 18  Region_gb               11567 non-null  uint8
 19  Region_ict              11567 non-null  uint8
 20  Region_ajk              11567 non-null  uint8
 21  Region_fata             11567 non-null  uint8
dtypes: float64(9), int32(2), int8(3), object(1), uint8(7)
memory usage: 1.2+ MB
```

[64]:
```
corr_diarrhea = df1.corr()['Diarrhea Occurrence'].drop('Diarrhea Occurrence').
 ↪to_frame()
corr_diarrhea
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_8076\4068845492.py:1: FutureWarning:
The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
  corr_diarrhea = df1.corr()['Diarrhea Occurrence'].drop('Diarrhea
Occurrence').to_frame()
```

[64]:
```
                         Diarrhea Occurrence
Age in Months                      -0.157505
mother's age                       -0.058772
```

```
No. Household Members              0.005120
Educational Attainment             0.003007
Mother's Working Status           -0.017547
improved_toilet                    0.009696
improved_water                    -0.010120
Gender_male                        0.017488
Size_Child                        -0.026767
Birth Weight                       0.010766
Age Difference Parents             0.022296
Wealth                            -0.006505
Residence_Urban                   -0.000938
Region_sindh                      -0.051487
Region_kpk                         0.032655
Region_balochistan                 0.004418
Region_gb                         -0.029760
Region_ict                         0.010295
Region_ajk                        -0.022873
Region_fata                        0.017391
```

[68]:
```
df1['Age in Months'].describe()
```

[68]:
```
count    11567.000000
mean        29.425175
std         17.502652
min          0.000000
25%         14.000000
50%         29.000000
75%         45.000000
max         59.000000
Name: Age in Months, dtype: float64
```

[147]:
```
df1.corr()
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_8076\473017434.py:1: FutureWarning:
The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
  df1.corr()
```

[147]:
```
                         Diarrhea Occurrence  Age in Months  mother's age  \
Diarrhea Occurrence                 1.000000      -0.157505     -0.058772
Age in Months                      -0.157505       1.000000      0.206964
mother's age                       -0.058772       0.206964      1.000000
No. Household Members               0.005120      -0.031891      0.009607
Educational Attainment              0.003007      -0.039141     -0.042848
Mother's Working Status            -0.017547       0.046624      0.082732
improved_toilet                     0.009696      -0.015354     -0.028302
improved_water                     -0.010120       0.008211     -0.010323
```

```
Gender_male                         0.017488       0.013389     -0.003060
Size_Child                         -0.026767       0.070107      0.018332
Birth Weight                        0.010766       0.018195      0.007014
Age Difference Parents              0.022296       0.011057     -0.038108
Wealth                             -0.006505      -0.023715     -0.031201
Residence_Urban                    -0.000938      -0.002430     -0.018633
Region_sindh                       -0.051487       0.002531     -0.010951
Region_kpk                          0.032655      -0.000109     -0.036881
Region_balochistan                  0.004418       0.008379      0.023179
Region_gb                          -0.029760       0.011715      0.037364
Region_ict                          0.010295      -0.012954     -0.001763
Region_ajk                         -0.022873      -0.005356      0.048594
Region_fata                         0.017391       0.002255     -0.034751

                         No. Household Members  Educational Attainment  \
Diarrhea Occurrence                   0.005120                0.003007
Age in Months                        -0.031891               -0.039141
mother's age                          0.009607               -0.042848
No. Household Members                 1.000000               -0.107459
Educational Attainment               -0.107459                1.000000
Mother's Working Status              -0.070182                0.031649
improved_toilet                       0.018685                0.300906
improved_water                       -0.003849                0.130701
Gender_male                          -0.013995                0.004288
Size_Child                            0.022745                0.053326
Birth Weight                          0.008320                0.023137
Age Difference Parents               -0.041280               -0.048365
Wealth                                0.059377                0.606153
Residence_Urban                      -0.001711                0.296724
Region_sindh                         -0.014175               -0.051142
Region_kpk                            0.102800               -0.051618
Region_balochistan                    0.119019               -0.170551
Region_gb                            -0.005928                0.050854
Region_ict                           -0.072634                0.150933
Region_ajk                           -0.102825                0.159184
Region_fata                           0.085197               -0.198047

                         Mother's Working Status  improved_toilet  \
Diarrhea Occurrence                    -0.017547         0.009696
Age in Months                           0.046624        -0.015354
mother's age                            0.082732        -0.028302
No. Household Members                  -0.070182         0.018685
Educational Attainment                  0.031649         0.300906
Mother's Working Status                 1.000000        -0.037131
improved_toilet                        -0.037131         1.000000
improved_water                          0.028749         0.143986
Gender_male                            -0.010876        -0.012939
```

```
Size_Child                             -0.028113         0.030048
Birth Weight                           -0.029152         0.021919
Age Difference Parents                 -0.015514        -0.001180
Wealth                                 -0.068047         0.495840
Residence_Urban                        -0.035112         0.288648
Region_sindh                            0.099947         0.029751
Region_kpk                             -0.088506         0.145145
Region_balochistan                     -0.001312        -0.126997
Region_gb                              -0.033674         0.022606
Region_ict                             -0.000305         0.102609
Region_ajk                             -0.006737        -0.127233
Region_fata                            -0.091206        -0.200747

                         improved_water  Gender_male  Size_Child  …  \
Diarrhea Occurrence           -0.010120     0.017488   -0.026767  …
Age in Months                  0.008211     0.013389    0.070107  …
mother's age                  -0.010323    -0.003060    0.018332  …
No. Household Members         -0.003849    -0.013995    0.022745  …
Educational Attainment         0.130701     0.004288    0.053326  …
Mother's Working Status        0.028749    -0.010876   -0.028113  …
improved_toilet                0.143986    -0.012939    0.030048  …
improved_water                 1.000000    -0.006249    0.024379  …
Gender_male                   -0.006249     1.000000   -0.006835  …
Size_Child                     0.024379    -0.006835    1.000000  …
Birth Weight                   0.001571     0.018343    0.199461  …
Age Difference Parents         0.003534     0.019028   -0.020508  …
Wealth                         0.242408     0.005159    0.077771  …
Residence_Urban                0.133856     0.009595    0.062290  …
Region_sindh                   0.074016    -0.018481    0.025457  …
Region_kpk                     0.010941    -0.006761    0.004446  …
Region_balochistan            -0.070838     0.008114    0.004155  …
Region_gb                     -0.030884     0.006842   -0.005124  …
Region_ict                     0.030028     0.001839    0.010979  …
Region_ajk                    -0.123982     0.013199   -0.006016  …
Region_fata                   -0.144786     0.007339    0.017201  …

                         Age Difference Parents    Wealth  Residence_Urban  \
Diarrhea Occurrence                    0.022296 -0.006505        -0.000938
Age in Months                          0.011057 -0.023715        -0.002430
mother's age                          -0.038108 -0.031201        -0.018633
No. Household Members                 -0.041280  0.059377        -0.001711
Educational Attainment                -0.048365  0.606153         0.296724
Mother's Working Status               -0.015514 -0.068047        -0.035112
improved_toilet                       -0.001180  0.495840         0.288648
improved_water                         0.003534  0.242408         0.133856
Gender_male                            0.019028  0.005159         0.009595
Size_Child                            -0.020508  0.077771         0.062290
```

```
Birth Weight                     -0.000714  0.025995         0.029884
Age Difference Parents            1.000000 -0.014707         0.019258
Wealth                           -0.014707  1.000000         0.503456
Residence_Urban                   0.019258  0.503456         1.000000
Region_sindh                     -0.034846 -0.077073         0.050316
Region_kpk                        0.098537  0.103998         0.051630
Region_balochistan               -0.006010 -0.130161         0.053752
Region_gb                        -0.001142 -0.132897        -0.075300
Region_ict                       -0.010562  0.231671         0.146028
Region_ajk                       -0.017517  0.029440         0.011782
Region_fata                       0.003101 -0.199906        -0.139509

                         Region_sindh  Region_kpk  Region_balochistan  \
Diarrhea Occurrence         -0.051487    0.032655            0.004418
Age in Months                0.002531   -0.000109            0.008379
mother's age                -0.010951   -0.036881            0.023179
No. Household Members       -0.014175    0.102800            0.119019
Educational Attainment      -0.051142   -0.051618           -0.170551
Mother's Working Status      0.099947   -0.088506           -0.001312
improved_toilet              0.029751    0.145145           -0.126997
improved_water               0.074016    0.010941           -0.070838
Gender_male                 -0.018481   -0.006761            0.008114
Size_Child                   0.025457    0.004446            0.004155
Birth Weight                -0.016701    0.010861           -0.003576
Age Difference Parents      -0.034846    0.098537           -0.006010
Wealth                      -0.077073    0.103998           -0.130161
Residence_Urban              0.050316    0.051630            0.053752
Region_sindh                 1.000000   -0.210963           -0.171047
Region_kpk                  -0.210963    1.000000           -0.164198
Region_balochistan          -0.171047   -0.164198            1.000000
Region_gb                   -0.130339   -0.125120           -0.101446
Region_ict                  -0.122999   -0.118074           -0.095733
Region_ajk                  -0.157632   -0.151321           -0.122690
Region_fata                 -0.144133   -0.138362           -0.112183

                         Region_gb  Region_ict  Region_ajk  Region_fata
Diarrhea Occurrence      -0.029760    0.010295   -0.022873     0.017391
Age in Months             0.011715   -0.012954   -0.005356     0.002255
mother's age              0.037364   -0.001763    0.048594    -0.034751
No. Household Members    -0.005928   -0.072634   -0.102825     0.085197
Educational Attainment    0.050854    0.150933    0.159184    -0.198047
Mother's Working Status  -0.033674   -0.000305   -0.006737    -0.091206
improved_toilet           0.022606    0.102609   -0.127233    -0.200747
improved_water           -0.030884    0.030028   -0.123982    -0.144786
Gender_male               0.006842    0.001839    0.013199     0.007339
Size_Child               -0.005124    0.010979   -0.006016     0.017201
Birth Weight              0.023699    0.013961   -0.017238     0.020257
```

```
Age Difference Parents   -0.001142 -0.010562 -0.017517  0.003101
Wealth                   -0.132897  0.231671  0.029440 -0.199906
Residence_Urban          -0.075300  0.146028  0.011782 -0.139509
Region_sindh             -0.130339 -0.122999 -0.157632 -0.144133
Region_kpk               -0.125120 -0.118074 -0.151321 -0.138362
Region_balochistan       -0.101446 -0.095733 -0.122690 -0.112183
Region_gb                 1.000000 -0.072949 -0.093490 -0.085484
Region_ict               -0.072949  1.000000 -0.088225 -0.080670
Region_ajk               -0.093490 -0.088225  1.000000 -0.103385
Region_fata              -0.085484 -0.080670 -0.103385  1.000000

[21 rows x 21 columns]
```

[ ]: