task6

October 18, 2024

[5]:
```
import pandas as pd
```

[6]:
```
data = pd.read_csv('C:\\Users\\SRINIVASA SESHADRI␣
↪K\\OneDrive\\Documents\\MainFlow Services and Technologies Internship\\Task␣
↪6 (Diabetes Analysis)\\diabetes_data.csv')
```

[7]:
```
type(data)
```

[7]:
```
pandas.core.frame.DataFrame
```

[8]:
```
data.shape
```

[8]:
```
(1879, 46)
```

[9]:
```
data.head(5)
```

[9]:

	PatientID	Age	Gender	Ethnicity	SocioeconomicStatus	EducationLevel	\
0	6000	44	0	1	2	1	
1	6001	51	1	0	1	2	
2	6002	89	1	0	1	3	
3	6003	21	1	1	1	2	
4	6004	27	1	0	1	3	

	BMI	Smoking	AlcoholConsumption	PhysicalActivity	…	\
0	32.985284	1	4.499365	2.443385	…	
1	39.916764	0	1.578919	8.301264	…	
2	19.782251	0	1.177301	6.103395	…	
3	32.376881	1	1.714621	8.645465	…	
4	16.808600	0	15.462549	4.629383	…	

	TinglingHandsFeet	QualityOfLifeScore	HeavyMetalsExposure	\
0	1	73.765109	0	
1	0	91.445753	0	
2	0	54.485744	0	
3	0	77.866758	0	
4	0	37.731808	0	

```
   OccupationalExposureChemicals  WaterQuality  MedicalCheckupsFrequency  \
0                              0             0                  1.782724
1                              0             1                  3.381070
2                              0             0                  2.701019
3                              0             1                  1.409056
4                              0             0                  1.218452

   MedicationAdherence  HealthLiteracy  Diagnosis  DoctorInCharge
0             4.486980        7.211349          1    Confidential
1             5.961705        5.024612          1    Confidential
2             8.950821        7.034944          0    Confidential
3             3.124769        4.717774          0    Confidential
4             6.977741        7.887940          0    Confidential

[5 rows x 46 columns]
```

[10]:
```
data.tail(4)
```

[10]:
```
      PatientID  Age  Gender  Ethnicity  SocioeconomicStatus  EducationLevel  \
1875       7875   80       1          0                    2               2
1876       7876   38       1          0                    0               2
1877       7877   43       0          1                    2               0
1878       7878   85       1          0                    2               2

            BMI  Smoking  AlcoholConsumption  PhysicalActivity  …  \
1875  27.694312        0           16.067905          7.107335  …
1876  35.640824        0            4.865124          9.881212  …
1877  32.423016        0            6.362936          4.750079  …
1878  33.145119        0           13.854861          5.434137  …

      TinglingHandsFeet  QualityOfLifeScore  HeavyMetalsExposure  \
1875                  0           77.128599                    0
1876                  0           13.148221                    0
1877                  0           54.370980                    0
1878                  1           43.720860                    0

      OccupationalExposureChemicals  WaterQuality  MedicalCheckupsFrequency  \
1875                              0             1                  0.424893
1876                              0             0                  0.553757
1877                              0             0                  1.132470
1878                              0             1                  3.070583

      MedicationAdherence  HealthLiteracy  Diagnosis  DoctorInCharge
1875             5.217465        0.915878          1    Confidential
1876             3.377744        3.017481          1    Confidential
1877             0.009250        4.914556          1    Confidential
1878             8.483128        7.790921          1    Confidential
```

```
[4 rows x 46 columns]
```

[11]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1879 entries, 0 to 1878
Data columns (total 46 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   PatientID                       1879 non-null   int64
 1   Age                             1879 non-null   int64
 2   Gender                          1879 non-null   int64
 3   Ethnicity                       1879 non-null   int64
 4   SocioeconomicStatus             1879 non-null   int64
 5   EducationLevel                  1879 non-null   int64
 6   BMI                             1879 non-null   float64
 7   Smoking                         1879 non-null   int64
 8   AlcoholConsumption              1879 non-null   float64
 9   PhysicalActivity                1879 non-null   float64
 10  DietQuality                     1879 non-null   float64
 11  SleepQuality                    1879 non-null   float64
 12  FamilyHistoryDiabetes           1879 non-null   int64
 13  GestationalDiabetes             1879 non-null   int64
 14  PolycysticOvarySyndrome         1879 non-null   int64
 15  PreviousPreDiabetes             1879 non-null   int64
 16  Hypertension                    1879 non-null   int64
 17  SystolicBP                      1879 non-null   int64
 18  DiastolicBP                     1879 non-null   int64
 19  FastingBloodSugar               1879 non-null   float64
 20  HbA1c                           1879 non-null   float64
 21  SerumCreatinine                 1879 non-null   float64
 22  BUNLevels                       1879 non-null   float64
 23  CholesterolTotal                1879 non-null   float64
 24  CholesterolLDL                  1879 non-null   float64
 25  CholesterolHDL                  1879 non-null   float64
 26  CholesterolTriglycerides        1879 non-null   float64
 27  AntihypertensiveMedications     1879 non-null   int64
 28  Statins                         1879 non-null   int64
 29  AntidiabeticMedications         1879 non-null   int64
 30  FrequentUrination               1879 non-null   int64
 31  ExcessiveThirst                 1879 non-null   int64
 32  UnexplainedWeightLoss           1879 non-null   int64
 33  FatigueLevels                   1879 non-null   float64
 34  BlurredVision                   1879 non-null   int64
 35  SlowHealingSores                1879 non-null   int64
 36  TinglingHandsFeet               1879 non-null   int64
```

```
 37  QualityOfLifeScore              1879 non-null   float64
 38  HeavyMetalsExposure             1879 non-null   int64
 39  OccupationalExposureChemicals   1879 non-null   int64
 40  WaterQuality                    1879 non-null   int64
 41  MedicalCheckupsFrequency        1879 non-null   float64
 42  MedicationAdherence             1879 non-null   float64
 43  HealthLiteracy                  1879 non-null   float64
 44  Diagnosis                       1879 non-null   int64
 45  DoctorInCharge                  1879 non-null   object
dtypes: float64(18), int64(27), object(1)
memory usage: 675.4+ KB
```

[12]:
```
data.describe()
```

[12]:

	PatientID	Age	Gender	Ethnicity	\
count	1879.000000	1879.000000	1879.000000	1879.000000	
mean	6939.000000	55.043108	0.487493	0.755721	
std	542.564896	20.515839	0.499977	1.047558	
min	6000.000000	20.000000	0.000000	0.000000	
25%	6469.500000	38.000000	0.000000	0.000000	
50%	6939.000000	55.000000	0.000000	0.000000	
75%	7408.500000	73.000000	1.000000	1.000000	
max	7878.000000	90.000000	1.000000	3.000000	

	SocioeconomicStatus	EducationLevel	BMI	Smoking	\
count	1879.000000	1879.000000	1879.000000	1879.000000	
mean	0.992017	1.699308	27.687601	0.281533	
std	0.764940	0.885665	7.190975	0.449866	
min	0.000000	0.000000	15.025898	0.000000	
25%	0.000000	1.000000	21.469981	0.000000	
50%	1.000000	2.000000	27.722988	0.000000	
75%	2.000000	2.000000	33.856460	1.000000	
max	2.000000	3.000000	39.998811	1.000000	

	AlcoholConsumption	PhysicalActivity	…	SlowHealingSores	\
count	1879.000000	1879.000000	…	1879.000000	
mean	10.096587	5.200790	…	0.102714	
std	5.914216	2.857012	…	0.303666	
min	0.000928	0.004089	…	0.000000	
25%	4.789725	2.751022	…	0.000000	
50%	10.173865	5.249002	…	0.000000	
75%	15.285359	7.671402	…	0.000000	
max	19.996231	9.993893	…	1.000000	

	TinglingHandsFeet	QualityOfLifeScore	HeavyMetalsExposure	\
count	1879.000000	1879.000000	1879.000000	
mean	0.111229	48.508643	0.052155	

```
std             0.314500          28.758488            0.222400
min             0.000000           0.002390            0.000000
25%             0.000000          23.974098            0.000000
50%             0.000000          47.519693            0.000000
75%             0.000000          72.883179            0.000000
max             1.000000          99.788530            1.000000

       OccupationalExposureChemicals  WaterQuality  MedicalCheckupsFrequency  \
count                    1879.000000   1879.000000               1879.000000
mean                        0.103246      0.200639                  1.997101
std                         0.304361      0.400585                  1.122632
min                         0.000000      0.000000                  0.004013
25%                         0.000000      0.000000                  1.057801
50%                         0.000000      0.000000                  1.987170
75%                         0.000000      0.000000                  2.946019
max                         1.000000      1.000000                  3.999715

       MedicationAdherence  HealthLiteracy    Diagnosis
count          1879.000000     1879.000000  1879.000000
mean              4.957539        5.011736     0.400213
std               2.910934        2.920908     0.490072
min               0.005384        0.000362     0.000000
25%               2.420024        2.410113     0.000000
50%               4.843886        5.035208     0.000000
75%               7.513933        7.586865     1.000000
max               9.997165        9.993029     1.000000

[8 rows x 45 columns]
```

[13]:
```
data = data.drop_duplicates()
data
```

[13]:
```
      PatientID  Age  Gender  Ethnicity  SocioeconomicStatus  EducationLevel  \
0          6000   44       0          1                    2               1
1          6001   51       1          0                    1               2
2          6002   89       1          0                    1               3
3          6003   21       1          1                    1               2
4          6004   27       1          0                    1               3
…           …  …     …        …                  …             …
1874       7874   37       0          0                    2               2
1875       7875   80       1          0                    2               2
1876       7876   38       1          0                    0               2
1877       7877   43       0          1                    2               0
1878       7878   85       1          0                    2               2

            BMI  Smoking  AlcoholConsumption  PhysicalActivity  …  \
0     32.985284        1            4.499365          2.443385  …
```

```
1     39.916764        0            1.578919          8.301264  …
2     19.782251        0            1.177301          6.103395  …
3     32.376881        1            1.714621          8.645465  …
4     16.808600        0           15.462549          4.629383  …
…         …       …                   …                …  …
1874  20.811137        0           10.946207          3.217636  …
1875  27.694312        0           16.067905          7.107335  …
1876  35.640824        0            4.865124          9.881212  …
1877  32.423016        0            6.362936          4.750079  …
1878  33.145119        0           13.854861          5.434137  …

      TinglingHandsFeet  QualityOfLifeScore  HeavyMetalsExposure  \
0                     1           73.765109                    0
1                     0           91.445753                    0
2                     0           54.485744                    0
3                     0           77.866758                    0
4                     0           37.731808                    0
…                   …                   …                    …
1874                  1           88.122729                    0
1875                  0           77.128599                    0
1876                  0           13.148221                    0
1877                  0           54.370980                    0
1878                  1           43.720860                    0

      OccupationalExposureChemicals  WaterQuality  MedicalCheckupsFrequency  \
0                                 0             0                  1.782724
1                                 0             1                  3.381070
2                                 0             0                  2.701019
3                                 0             1                  1.409056
4                                 0             0                  1.218452
…                               …             …                         …
1874                              0             1                  3.154225
1875                              0             1                  0.424893
1876                              0             0                  0.553757
1877                              0             0                  1.132470
1878                              0             1                  3.070583

      MedicationAdherence  HealthLiteracy  Diagnosis  DoctorInCharge
0                4.486980        7.211349          1    Confidential
1                5.961705        5.024612          1    Confidential
2                8.950821        7.034944          0    Confidential
3                3.124769        4.717774          0    Confidential
4                6.977741        7.887940          0    Confidential
…                   …             …        …            …
1874             3.849584        8.805087          0    Confidential
1875             5.217465        0.915878          1    Confidential
1876             3.377744        3.017481          1    Confidential
```

```
1877            0.009250        4.914556          1    Confidential
1878            8.483128        7.790921          1    Confidential

[1879 rows x 46 columns]
```

[14]:
```
data.isnull()
```

[14]:
```
      PatientID    Age  Gender  Ethnicity  SocioeconomicStatus  \
0         False  False   False      False                False
1         False  False   False      False                False
2         False  False   False      False                False
3         False  False   False      False                False
4         False  False   False      False                False
…           …    …       …          …                    …
1874      False  False   False      False                False
1875      False  False   False      False                False
1876      False  False   False      False                False
1877      False  False   False      False                False
1878      False  False   False      False                False

      EducationLevel    BMI  Smoking  AlcoholConsumption  PhysicalActivity  \
0              False  False    False               False             False
1              False  False    False               False             False
2              False  False    False               False             False
3              False  False    False               False             False
4              False  False    False               False             False
…                …    …        …                   …                 …
1874           False  False    False               False             False
1875           False  False    False               False             False
1876           False  False    False               False             False
1877           False  False    False               False             False
1878           False  False    False               False             False

      …  TinglingHandsFeet  QualityOfLifeScore  HeavyMetalsExposure  \
0     …              False               False                False
1     …              False               False                False
2     …              False               False                False
3     …              False               False                False
4     …              False               False                False
…   …                …                   …                    …
1874  …              False               False                False
1875  …              False               False                False
1876  …              False               False                False
1877  …              False               False                False
1878  …              False               False                False

      OccupationalExposureChemicals  WaterQuality  MedicalCheckupsFrequency  \
```

```
0                          False         False                    False
1                          False         False                    False
2                          False         False                    False
3                          False         False                    False
4                          False         False                    False
…                            …             …                        …
1874                       False         False                    False
1875                       False         False                    False
1876                       False         False                    False
1877                       False         False                    False
1878                       False         False                    False

      MedicationAdherence  HealthLiteracy  Diagnosis  DoctorInCharge
0                   False           False      False           False
1                   False           False      False           False
2                   False           False      False           False
3                   False           False      False           False
4                   False           False      False           False
…                     …               …          …               …
1874                False           False      False           False
1875                False           False      False           False
1876                False           False      False           False
1877                False           False      False           False
1878                False           False      False           False

[1879 rows x 46 columns]
```

[15]:
```
data.isnull().sum()
```

[15]:
```
PatientID                  0
Age                        0
Gender                     0
Ethnicity                  0
SocioeconomicStatus        0
EducationLevel             0
BMI                        0
Smoking                    0
AlcoholConsumption         0
PhysicalActivity           0
DietQuality                0
SleepQuality               0
FamilyHistoryDiabetes      0
GestationalDiabetes        0
PolycysticOvarySyndrome    0
PreviousPreDiabetes        0
Hypertension               0
SystolicBP                 0
```

```
DiastolicBP                     0
FastingBloodSugar               0
HbA1c                           0
SerumCreatinine                 0
BUNLevels                       0
CholesterolTotal                0
CholesterolLDL                  0
CholesterolHDL                  0
CholesterolTriglycerides        0
AntihypertensiveMedications     0
Statins                         0
AntidiabeticMedications         0
FrequentUrination               0
ExcessiveThirst                 0
UnexplainedWeightLoss           0
FatigueLevels                   0
BlurredVision                   0
SlowHealingSores                0
TinglingHandsFeet               0
QualityOfLifeScore              0
HeavyMetalsExposure             0
OccupationalExposureChemicals   0
WaterQuality                    0
MedicalCheckupsFrequency        0
MedicationAdherence             0
HealthLiteracy                  0
Diagnosis                       0
DoctorInCharge                  0
dtype: int64
```

[16]:
```
data.isnull().sum().sum()
```

[16]:
```
0
```

[17]:
```
import numpy as np
from scipy import stats
```

[18]:
```
data.columns
```

[18]:
```
Index(['PatientID', 'Age', 'Gender', 'Ethnicity', 'SocioeconomicStatus',
       'EducationLevel', 'BMI', 'Smoking', 'AlcoholConsumption',
       'PhysicalActivity', 'DietQuality', 'SleepQuality',
       'FamilyHistoryDiabetes', 'GestationalDiabetes',
       'PolycysticOvarySyndrome', 'PreviousPreDiabetes', 'Hypertension',
       'SystolicBP', 'DiastolicBP', 'FastingBloodSugar', 'HbA1c',
       'SerumCreatinine', 'BUNLevels', 'CholesterolTotal', 'CholesterolLDL',
       'CholesterolHDL', 'CholesterolTriglycerides',
```

```
       'AntihypertensiveMedications', 'Statins', 'AntidiabeticMedications',
       'FrequentUrination', 'ExcessiveThirst', 'UnexplainedWeightLoss',
       'FatigueLevels', 'BlurredVision', 'SlowHealingSores',
       'TinglingHandsFeet', 'QualityOfLifeScore', 'HeavyMetalsExposure',
       'OccupationalExposureChemicals', 'WaterQuality',
       'MedicalCheckupsFrequency', 'MedicationAdherence', 'HealthLiteracy',
       'Diagnosis', 'DoctorInCharge'],
      dtype='object')
```

[19]:
```
data.drop(['PatientID', 'Gender', 'Ethnicity', 'SocioeconomicStatus',
           'EducationLevel', 'Smoking', 'AlcoholConsumption',
           'FamilyHistoryDiabetes', 'GestationalDiabetes',
           'PolycysticOvarySyndrome', 'PreviousPreDiabetes',
           'Hypertension', 'AntihypertensiveMedications',
           'Statins', 'AntidiabeticMedications', 'FrequentUrination',
           'ExcessiveThirst', 'UnexplainedWeightLoss',
           'BlurredVision', 'SlowHealingSores', 'TinglingHandsFeet',
           'HeavyMetalsExposure', 'OccupationalExposureChemicals',
           'WaterQuality', 'Diagnosis', 'DoctorInCharge'],
          axis=1, inplace=True)
print(data.head())
```

```
   Age        BMI  PhysicalActivity  DietQuality  SleepQuality  SystolicBP  \
0   44  32.985284          2.443385     4.898831      4.049885          93
1   51  39.916764          8.301264     8.941093      7.508150         165
2   89  19.782251          6.103395     7.722543      7.708387         119
3   21  32.376881          8.645465     4.804044      6.286548         169
4   27  16.808600          4.629383     2.532756      9.771125         165

   DiastolicBP  FastingBloodSugar     HbA1c  SerumCreatinine  BUNLevels  \
0           73         163.687162  9.283631         2.665607  28.190147
1           99         188.347070  7.326870         4.172177  32.149491
2           91         127.703653  4.083426         1.973168  10.018375
3           87          82.688415  6.516645         3.057797  44.123281
4           69          90.743395  5.607222         4.150353   7.757117

   CholesterolTotal  CholesterolLDL  CholesterolHDL  CholesterolTriglycerides  \
0        254.270670       86.993627       70.801469                190.335834
1        155.358831      110.056105       39.900112                 81.172469
2        231.608922       62.035793       62.480666                279.809069
3        176.592374       68.238410       46.977819                112.751396
4        157.344121       66.476215       40.059755                381.528785

   FatigueLevels  QualityOfLifeScore  MedicalCheckupsFrequency  \
0       9.534169           73.765109                  1.782724
1       0.123214           91.445753                  3.381070
2       9.643320           54.485744                  2.701019
```

```
3       3.403557           77.866758                  1.409056
4       2.924687           37.731808                  1.218452

   MedicationAdherence  HealthLiteracy
0             4.486980        7.211349
1             5.961705        5.024612
2             8.950821        7.034944
3             3.124769        4.717774
4             6.977741        7.887940
```

[20]:
```
Q1=data.quantile(0.25)
Q3=data.quantile(0.75)
IQR=Q3-Q1
print(IQR)
```

```
Age                          35.000000
BMI                          12.386479
PhysicalActivity              4.920380
DietQuality                   4.879256
SleepQuality                  3.042025
SystolicBP                   44.000000
DiastolicBP                  30.000000
FastingBloodSugar            65.020892
HbA1c                         2.979356
SerumCreatinine               2.243660
BUNLevels                    21.334289
CholesterolTotal             76.596373
CholesterolLDL               73.813498
CholesterolHDL               40.047150
CholesterolTriglycerides    172.535014
FatigueLevels                 5.152024
QualityOfLifeScore           48.909081
MedicalCheckupsFrequency      1.888217
MedicationAdherence           5.093910
HealthLiteracy                5.176752
dtype: float64
```

[21]:
```
data[~((data< (Q1-1.5*IQR) ) | (data>(Q3+1.5*IQR)) ) .any(axis=1)]
data
```

[21]:

	Age	BMI	PhysicalActivity	DietQuality	SleepQuality	SystolicBP	\
0	44	32.985284	2.443385	4.898831	4.049885	93	
1	51	39.916764	8.301264	8.941093	7.508150	165	
2	89	19.782251	6.103395	7.722543	7.708387	119	
3	21	32.376881	8.645465	4.804044	6.286548	169	
4	27	16.808600	4.629383	2.532756	9.771125	165	
…	…	…	…	…	…	…	

```
1874   37  20.811137          3.217636     8.338196      8.703430         104
1875   80  27.694312          7.107335     3.034771      4.472689         166
1876   38  35.640824          9.881212     2.657002      4.812610         128
1877   43  32.423016          4.750079     8.736024      7.017390         124
1878   85  33.145119          5.434137     5.127496      4.924963         134

      DiastolicBP  FastingBloodSugar     HbA1c  SerumCreatinine  BUNLevels  \
0              73         163.687162  9.283631         2.665607  28.190147
1              99         188.347070  7.326870         4.172177  32.149491
2              91         127.703653  4.083426         1.973168  10.018375
3              87          82.688415  6.516645         3.057797  44.123281
4              69          90.743395  5.607222         4.150353   7.757117
…             …                 …         …                …         …
1874           74         109.832032  5.920723         3.984707  21.645433
1875          115          90.729361  7.332397         2.132178   7.433835
1876           70         149.366801  4.907208         2.195365  26.225481
1877           91         162.027044  8.820613         0.893745  41.555665
1878           86         175.011749  7.814477         4.607711  28.471762

      CholesterolTotal  CholesterolLDL  CholesterolHDL  \
0           254.270670       86.993627       70.801469
1           155.358831      110.056105       39.900112
2           231.608922       62.035793       62.480666
3           176.592374       68.238410       46.977819
4           157.344121       66.476215       40.059755
…                  …              …               …
1874        260.342336       99.720234       40.296248
1875        273.728852      179.858432       48.873298
1876        293.513379      113.915759       62.217083
1877        178.559550      141.601955       74.116118
1878        268.635952       57.431715       73.728242

      CholesterolTriglycerides  FatigueLevels  QualityOfLifeScore  \
0                   190.335834       9.534169           73.765109
1                    81.172469       0.123214           91.445753
2                   279.809069       9.643320           54.485744
3                   112.751396       3.403557           77.866758
4                   381.528785       2.924687           37.731808
…                          …              …                   …
1874                198.613903       3.693506           88.122729
1875                271.239061       4.225031           77.128599
1876                374.429055       1.174257           13.148221
1877                171.298228       9.732583           54.370980
1878                174.869266       4.360088           43.720860

      MedicalCheckupsFrequency  MedicationAdherence  HealthLiteracy
0                     1.782724             4.486980        7.211349
```

```
1                    3.381070             5.961705        5.024612
2                    2.701019             8.950821        7.034944
3                    1.409056             3.124769        4.717774
4                    1.218452             6.977741        7.887940
…                         …                    …             …
1874                 3.154225             3.849584        8.805087
1875                 0.424893             5.217465        0.915878
1876                 0.553757             3.377744        3.017481
1877                 1.132470             0.009250        4.914556
1878                 3.070583             8.483128        7.790921

[1879 rows x 20 columns]
```

[22]:
```
data.describe()
```

[22]:
```
                Age          BMI  PhysicalActivity  DietQuality  SleepQuality  \
count  1879.000000  1879.000000       1879.000000  1879.000000   1879.000000
mean     55.043108    27.687601          5.200790     4.895801      7.021328
std      20.515839     7.190975          2.857012     2.867144      1.729469
min      20.000000    15.025898          0.004089     0.000885      4.004336
25%      38.000000    21.469981          2.751022     2.476802      5.481789
50%      55.000000    27.722988          5.249002     4.888566      7.094692
75%      73.000000    33.856460          7.671402     7.356058      8.523814
max      90.000000    39.998811          9.993893     9.998677      9.989372

        SystolicBP  DiastolicBP  FastingBloodSugar        HbA1c  \
count  1879.000000  1879.000000        1879.000000  1879.000000
mean    134.050559    89.863757         135.204490     6.976133
std      25.613830    17.328086          37.515750     1.739365
min      90.000000    60.000000          70.074649     4.003089
25%     112.000000    75.000000         102.341470     5.443856
50%     134.000000    90.000000         137.398241     7.095732
75%     156.000000   105.000000         167.362362     8.423211
max     179.000000   119.000000         199.935506     9.991193

       SerumCreatinine    BUNLevels  CholesterolTotal  CholesterolLDL  \
count      1879.000000  1879.000000       1879.000000     1879.000000
mean          2.784590    27.798153        225.006464      124.656831
std           1.308023    12.800797         43.367170       42.911145
min           0.500565     5.010401        150.056094       50.058252
25%           1.654472    17.172009        186.933051       87.810946
50%           2.855105    28.190147        225.120112      124.918023
75%           3.898133    38.506299        263.529424      161.624444
max           4.993974    49.975728        299.998480      199.898732

       CholesterolHDL  CholesterolTriglycerides  FatigueLevels  \
count     1879.000000               1879.000000    1879.000000
```

```
mean        60.060944                227.386167       4.949003
std         23.316682                101.071578       2.884483
min         20.014494                 50.154649       0.004977
25%         40.011963                140.873930       2.417748
50%         60.456988                228.417429       4.851914
75%         80.059112                313.408944       7.569772
max         99.958394                399.885928       9.999979

       QualityOfLifeScore  MedicalCheckupsFrequency  MedicationAdherence  \
count         1879.000000               1879.000000          1879.000000
mean            48.508643                  1.997101             4.957539
std             28.758488                  1.122632             2.910934
min              0.002390                  0.004013             0.005384
25%             23.974098                  1.057801             2.420024
50%             47.519693                  1.987170             4.843886
75%             72.883179                  2.946019             7.513933
max             99.788530                  3.999715             9.997165

       HealthLiteracy
count     1879.000000
mean         5.011736
std          2.920908
min          0.000362
25%          2.410113
50%          5.035208
75%          7.586865
max          9.993029
```

[23]:
```
print(data.columns)
```

```
Index(['Age', 'BMI', 'PhysicalActivity', 'DietQuality', 'SleepQuality',
       'SystolicBP', 'DiastolicBP', 'FastingBloodSugar', 'HbA1c',
       'SerumCreatinine', 'BUNLevels', 'CholesterolTotal', 'CholesterolLDL',
       'CholesterolHDL', 'CholesterolTriglycerides', 'FatigueLevels',
       'QualityOfLifeScore', 'MedicalCheckupsFrequency', 'MedicationAdherence',
       'HealthLiteracy'],
      dtype='object')
```

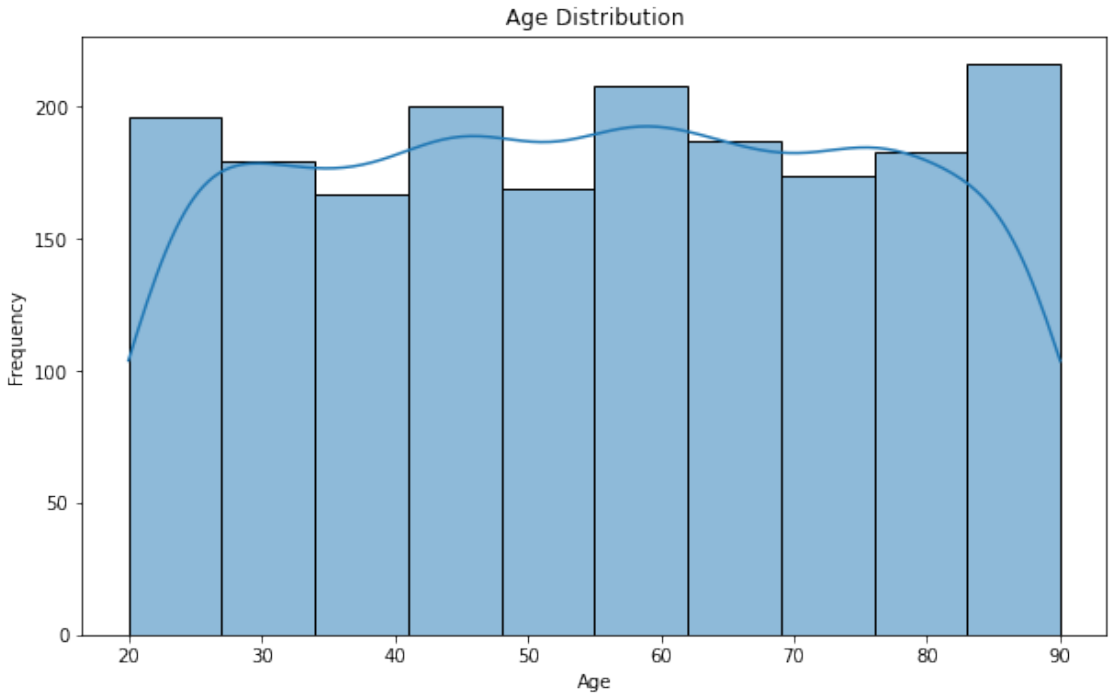
[67]:
```
'''
Questions
1. What is the overall distribution of ages in the dataset?
2. What is the distribution of BMI values among the participants?
3. How does physical activity level correlate with BMI?
4. Are there trends in physical activity levels across different age groups?
5. What are the average systolic and diastolic blood pressure readings across␣
 ↪different age groups?
6. Are there patterns in fasting blood sugar levels based on BMI and age?
```

```
7. How do total cholesterol, LDL, HDL, and triglycerides compare among␣
 ↪different age groups?
'''
```

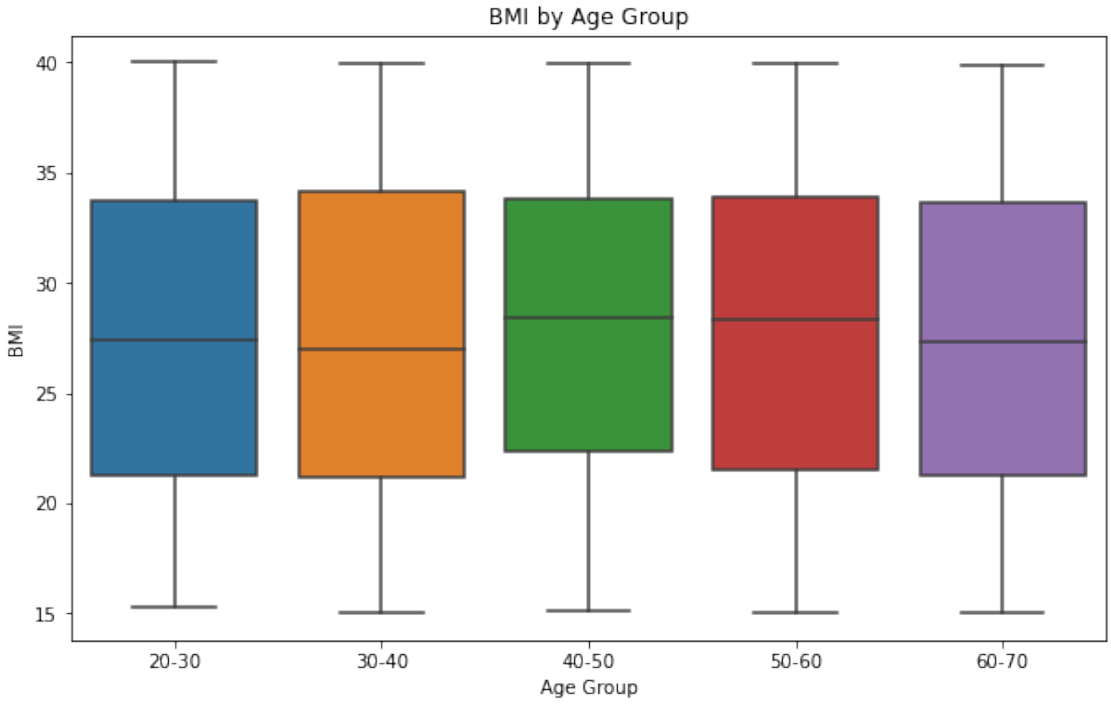
[67]:
```
'\nQuestions \n1. What is the overall distribution of ages in the dataset?\n2.
What is the distribution of BMI values among the participants?\n3. How does
physical activity level correlate with BMI?\n4. Are there trends in physical
activity levels across different age groups?\n5. What are the average systolic
and diastolic blood pressure readings across different age groups?\n6. Are there
patterns in fasting blood sugar levels based on BMI and age?\n7. How do total
cholesterol, LDL, HDL, and triglycerides compare among different age groups?\n'
```

[25]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

[41]:
```
# What is the overall distribution of ages in the dataset?
plt.figure(figsize=(10, 6))
sns.histplot(data['Age'], bins=10, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



[27]:
```
#  What is the distribution of BMI values among the participants?
data['AgeGroup'] = pd.cut(data['Age'], bins=[20, 30, 40, 50, 60, 70],␣
 ↪labels=['20-30', '30-40', '40-50', '50-60', '60-70'])
plt.figure(figsize=(10, 6))
sns.boxplot(data=data, x='AgeGroup', y='BMI')
plt.title('BMI by Age Group')
plt.xlabel('Age Group')
plt.ylabel('BMI')
plt.show()
```

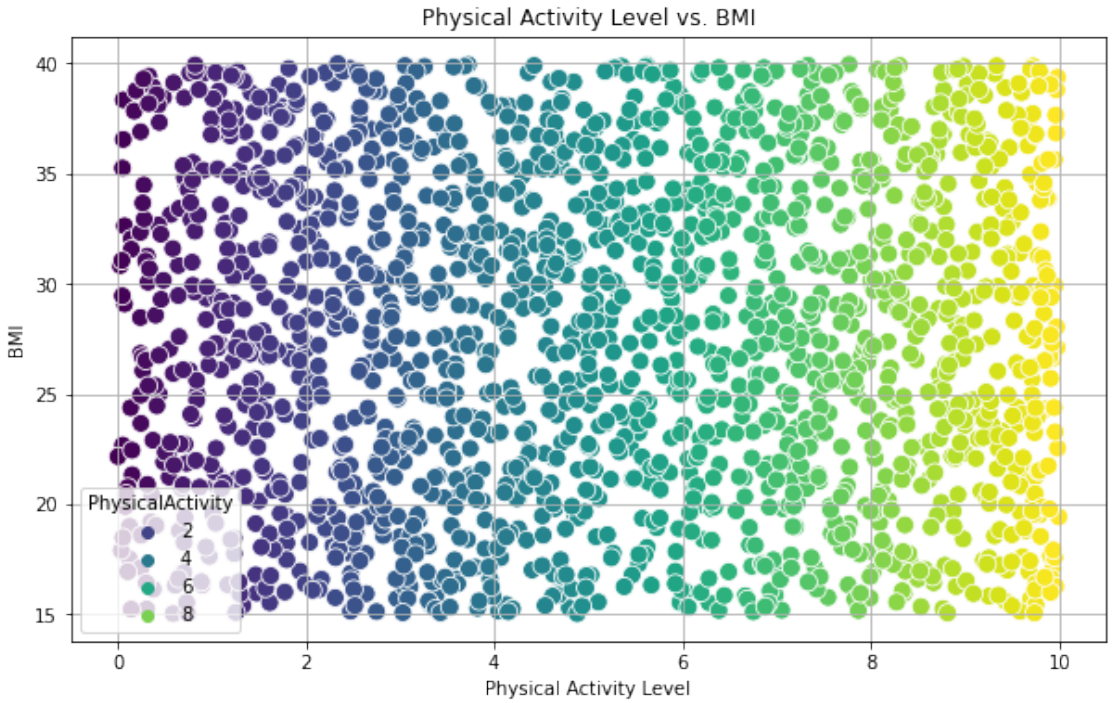


[28]:
```
print(data[['PhysicalActivity', 'BMI', 'SystolicBP', 'DiastolicBP']].describe())
print(data[['PhysicalActivity', 'BMI', 'SystolicBP', 'DiastolicBP']].nunique())
```

```
       PhysicalActivity          BMI   SystolicBP  DiastolicBP
count       1879.000000  1879.000000  1879.000000  1879.000000
mean           5.200790    27.687601   134.050559    89.863757
std            2.857012     7.190975    25.613830    17.328086
min            0.004089    15.025898    90.000000    60.000000
25%            2.751022    21.469981   112.000000    75.000000
50%            5.249002    27.722988   134.000000    90.000000
75%            7.671402    33.856460   156.000000   105.000000
max            9.993893    39.998811   179.000000   119.000000
PhysicalActivity    1879
BMI                 1879
```

```
SystolicBP            90
DiastolicBP           60
dtype: int64
```

[35]:
```
# How does physical activity level correlate with BMI or health metrics like␣
 ↪blood pressure?
# Scatter Plot for Physical Activity vs. BMI
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='PhysicalActivity', y='BMI',␣
 ↪hue='PhysicalActivity', palette='viridis', s=100)
plt.title('Physical Activity Level vs. BMI')
plt.xlabel('Physical Activity Level')
plt.ylabel('BMI')
plt.grid()
plt.show()
```



[30]:
```
# How does physical activity level correlate with BMI or health metrics like␣
 ↪blood pressure?
# Calculating and printing correlation coefficient for Physical Activity and BMI
correlation_bmi = data['PhysicalActivity'].corr(data['BMI'])
print(f'Correlation between Physical Activity and BMI: {correlation_bmi:.2f}')
```

```
Correlation between Physical Activity and BMI: -0.00
```

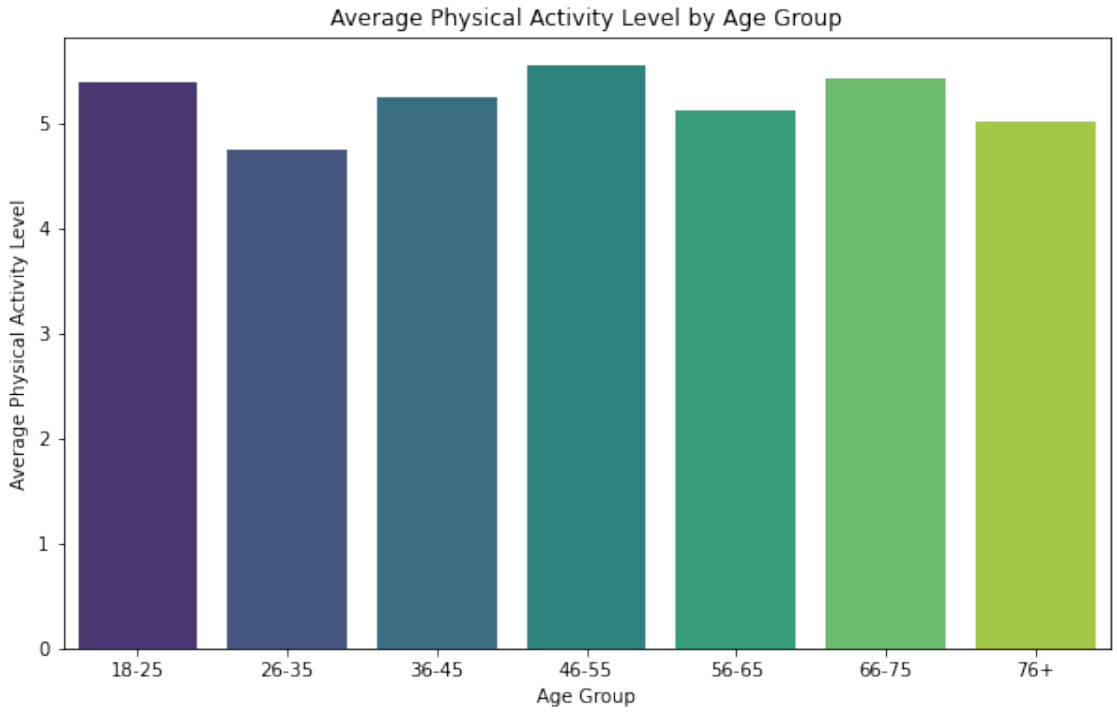
[43]:
```
pd.crosstab(data.PhysicalActivity, data.AgeGroup)
```

[43]:

AgeGroup	18-25	26-35	36-45	46-55	56-65	66-75	76+
PhysicalActivity							
0.004089	0	0	0	0	1	0	0
0.038327	0	0	1	0	0	0	0
0.045043	0	1	0	0	0	0	0
0.050365	0	1	0	0	0	0	0
0.051823	0	0	0	0	0	1	0
…	…	…	…	…	…	…	…
9.969412	0	0	0	0	1	0	0
9.969572	0	1	0	0	0	0	0
9.974534	0	1	0	0	0	0	0
9.980205	0	0	0	1	0	0	0
9.980646	0	0	0	1	0	0	0

[1714 rows x 7 columns]

[40]:
```
#Are there trends in physical activity levels across different age groups?
# Creating age groups
bins = [18, 25, 35, 45, 55, 65, 75, 85]
labels = ['18-25', '26-35', '36-45', '46-55', '56-65', '66-75', '76+']
data['AgeGroup'] = pd.cut(data['Age'], bins=bins, labels=labels, right=False)

# Calculating average physical activity level by age group
avg_activity_by_age = data.groupby('AgeGroup')['PhysicalActivity'].mean().
 ↪reset_index()

# Bar Plot for Average Physical Activity by Age Group
plt.figure(figsize=(10, 6))
sns.barplot(data=avg_activity_by_age, x='AgeGroup', y='PhysicalActivity',␣
 ↪palette='viridis')
plt.title('Average Physical Activity Level by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Average Physical Activity Level')
plt.show()
```



[73]:
```
pd.crosstab(data.AgeGroup, data.SystolicBP)
```

[73]:

SystolicBP	90	91	92	93	94	95	96	97	98	99	…	170	171	\
AgeGroup											…			
18-25	0	2	1	1	1	3	2	5	0	2	…	2	1	
26-35	6	4	2	3	8	1	1	2	3	6	…	7	3	
36-45	1	2	0	4	1	4	3	4	2	1	…	2	2	
46-55	3	1	1	5	5	5	4	1	2	2	…	5	3	
56-65	2	3	1	3	1	1	5	0	2	2	…	7	3	
66-75	2	1	0	5	3	3	5	4	3	2	…	1	1	
76+	1	4	4	3	5	6	4	3	4	6	…	6	2	

SystolicBP	172	173	174	175	176	177	178	179
AgeGroup								
18-25	2	1	1	0	0	0	1	4
26-35	2	2	2	1	2	2	5	4
36-45	3	3	3	0	2	1	2	2
46-55	3	6	4	3	1	0	2	4
56-65	5	5	5	2	3	1	7	3
66-75	2	1	1	1	3	5	4	1
76+	2	4	4	4	1	2	5	2

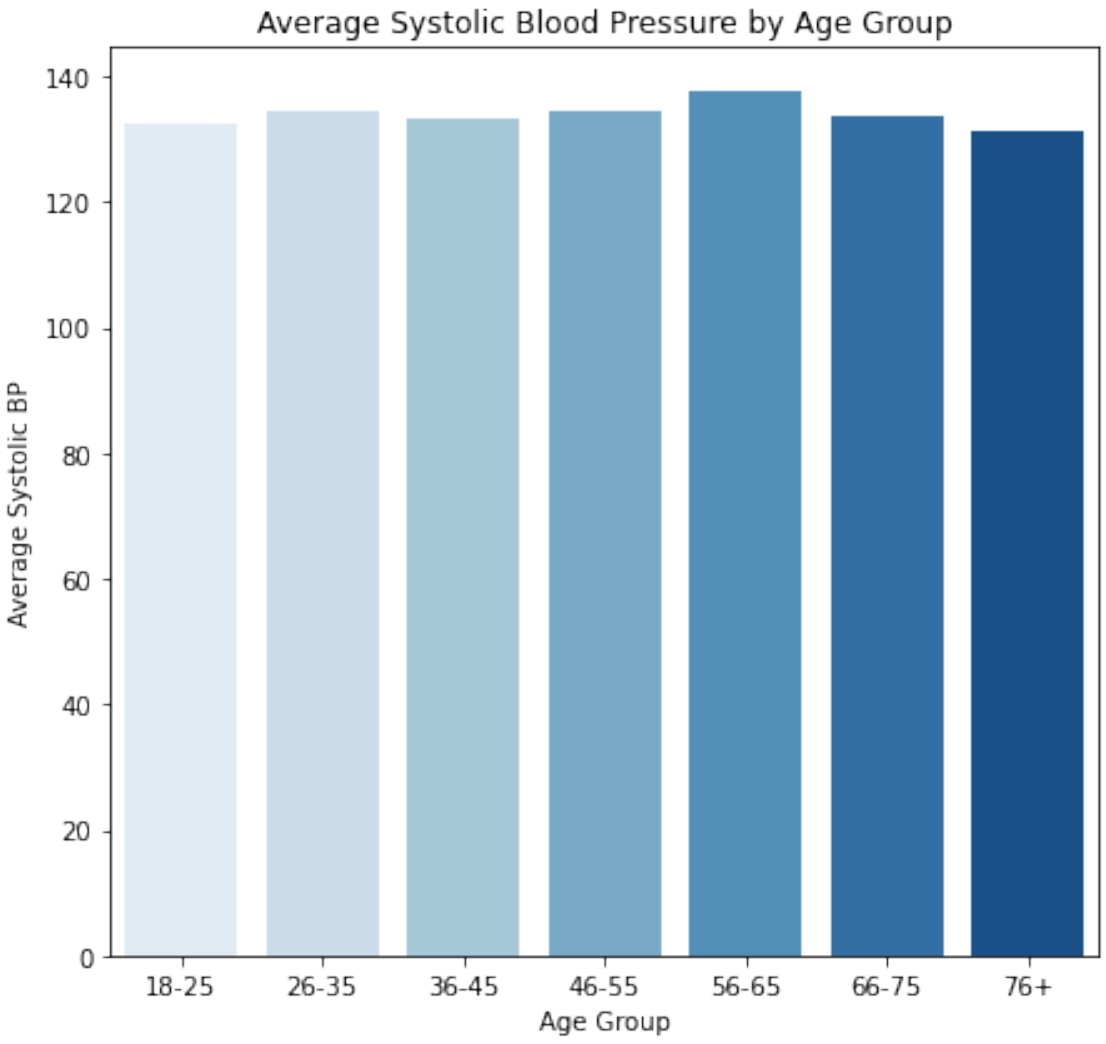
[7 rows x 90 columns]

[70]:
```
pd.crosstab(data.AgeGroup, data.DiastolicBP)
```

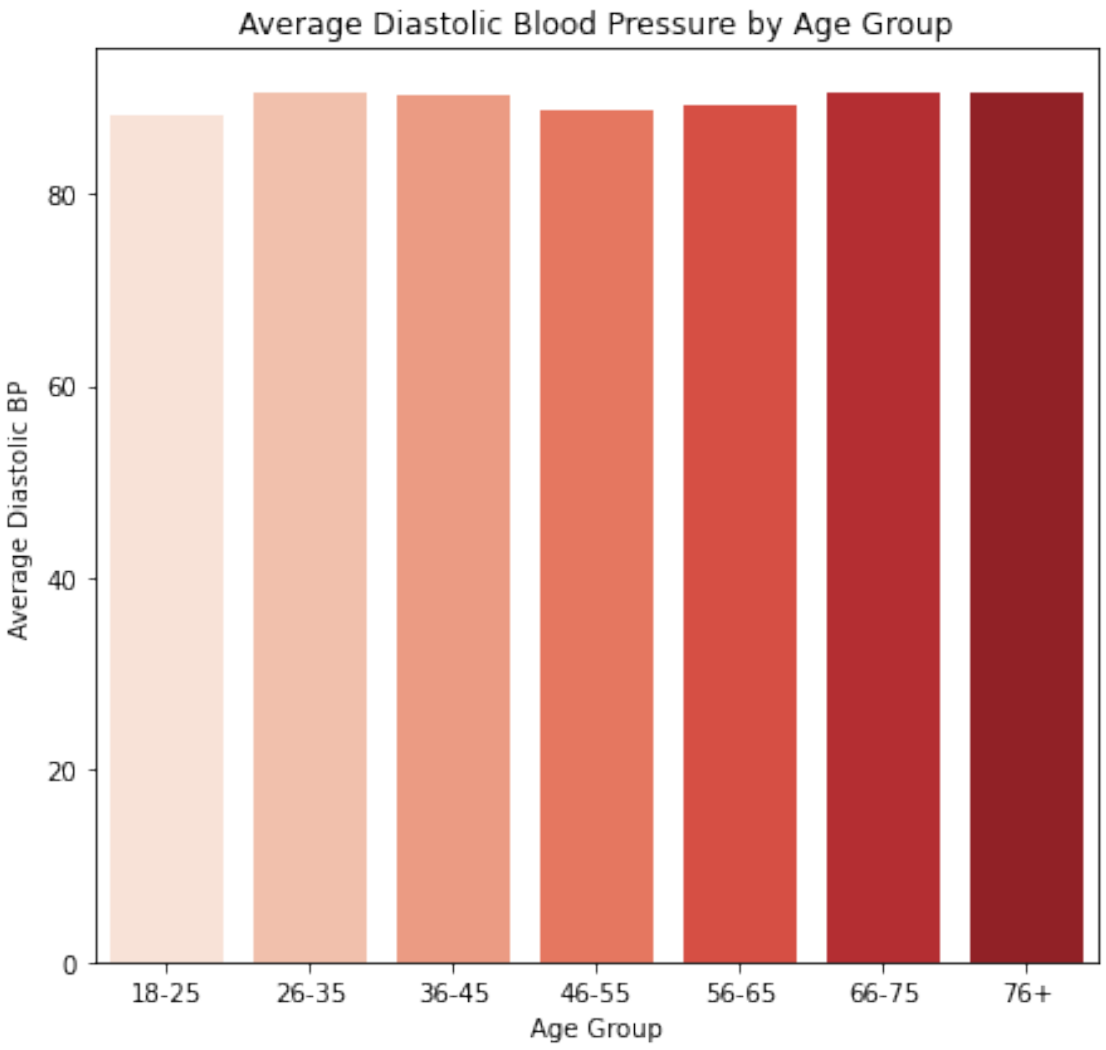
[70]:
```
DiastolicBP  60   61   62   63   64   65   66   67   68   69   …  110  111  \
AgeGroup                                                        …
18-25         1    2    2    1    4    6    3    5    1    2  …    1    2
26-35         5    4    8    2    3    2    2    1    5    4  …    6    3
36-45         4    5    2    1    8    4    4    3    6    4  …    5    4
46-55         4    2    1    4    9    6    5    4    9    3  …    1    2
56-65         6    3    8    4    6    1    3    5    8    5  …    7    9
66-75         3    4    4   11    3    4    5    4    2    4  …    4    9
76+           2    2    1    4    7    7    1    2    3    6  …    5    4

DiastolicBP  112  113  114  115  116  117  118  119
AgeGroup
18-25         2    1    4    3    0    2    2    4
26-35         4    5    5    7    5    1    4   10
36-45         3    9    7    3    5    3    5    5
46-55         4    4    2    4    4    1    6    4
56-65         6    5    4    5    2    7    4    3
66-75        11    1    5    5    3    5    4    5
76+           3    3    4    2    5    5    9    5

[7 rows x 60 columns]
```

[45]:
```
# What are the average systolic and diastolic blood pressure readings across␣
 ↪different age groups?
# Creating age groups
bins = [18, 25, 35, 45, 55, 65, 75, 85]
labels = ['18-25', '26-35', '36-45', '46-55', '56-65', '66-75', '76+']
data['AgeGroup'] = pd.cut(data['Age'], bins=bins, labels=labels, right=False)

# Calculating average systolic and diastolic blood pressure by age group
avg_bp_by_age = data.groupby('AgeGroup')[['SystolicBP', 'DiastolicBP']].mean().
 ↪reset_index()
```

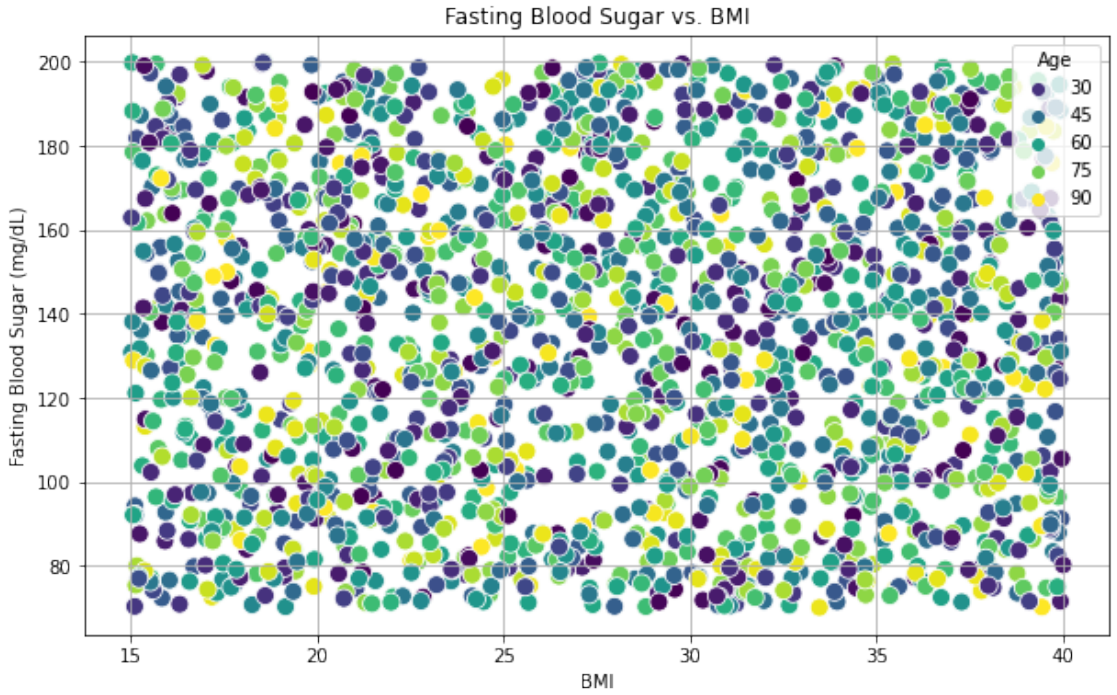
[50]:
```
# Bar Plot for Average Systolic Blood Pressure
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.barplot(data=avg_bp_by_age, x='AgeGroup', y='SystolicBP', palette='Blues')
plt.title('Average Systolic Blood Pressure by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Average Systolic BP')
plt.tight_layout()
plt.show()
```



[52]:
```
# Bar Plot for Average Diastolic Blood Pressure
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 2)
sns.barplot(data=avg_bp_by_age, x='AgeGroup', y='DiastolicBP', palette='Reds')
plt.title('Average Diastolic Blood Pressure by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Average Diastolic BP')
plt.tight_layout()
plt.show()
```



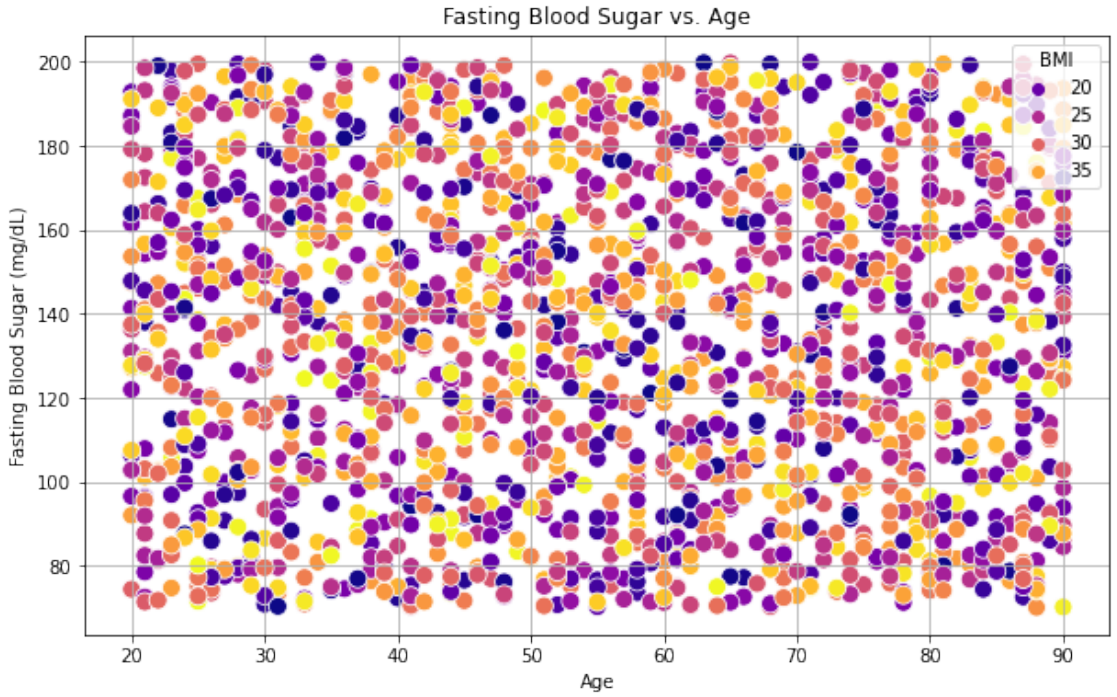
[53]:
```
#Are there patterns in fasting blood sugar levels based on BMI and age?
# Scatter Plot for Fasting Blood Sugar vs. BMI
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='BMI', y='FastingBloodSugar', hue='Age',␣
 ↪palette='viridis', s=100)
plt.title('Fasting Blood Sugar vs. BMI')
plt.xlabel('BMI')
plt.ylabel('Fasting Blood Sugar (mg/dL)')
plt.grid()
plt.show()
```



[54]:
```
correlation_bmi = data['FastingBloodSugar'].corr(data['BMI'])
print(f'Correlation between Fasting Blood Sugar and BMI: {correlation_bmi:.2f}')
```

```
Correlation between Fasting Blood Sugar and BMI: -0.01
```

[56]:
```
#Are there patterns in fasting blood sugar levels based on BMI and age?
# Scatter Plot for Fasting Blood Sugar vs. Age
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='Age', y='FastingBloodSugar', hue='BMI',␣
 ↪palette='plasma', s=100)
plt.title('Fasting Blood Sugar vs. Age')
plt.xlabel('Age')
plt.ylabel('Fasting Blood Sugar (mg/dL)')
plt.grid()
plt.show()
```



[57]:
```
correlation_age = data['FastingBloodSugar'].corr(data['Age'])
print(f'Correlation between Fasting Blood Sugar and Age: {correlation_age:.2f}')
```

```
Correlation between Fasting Blood Sugar and Age: -0.02
```

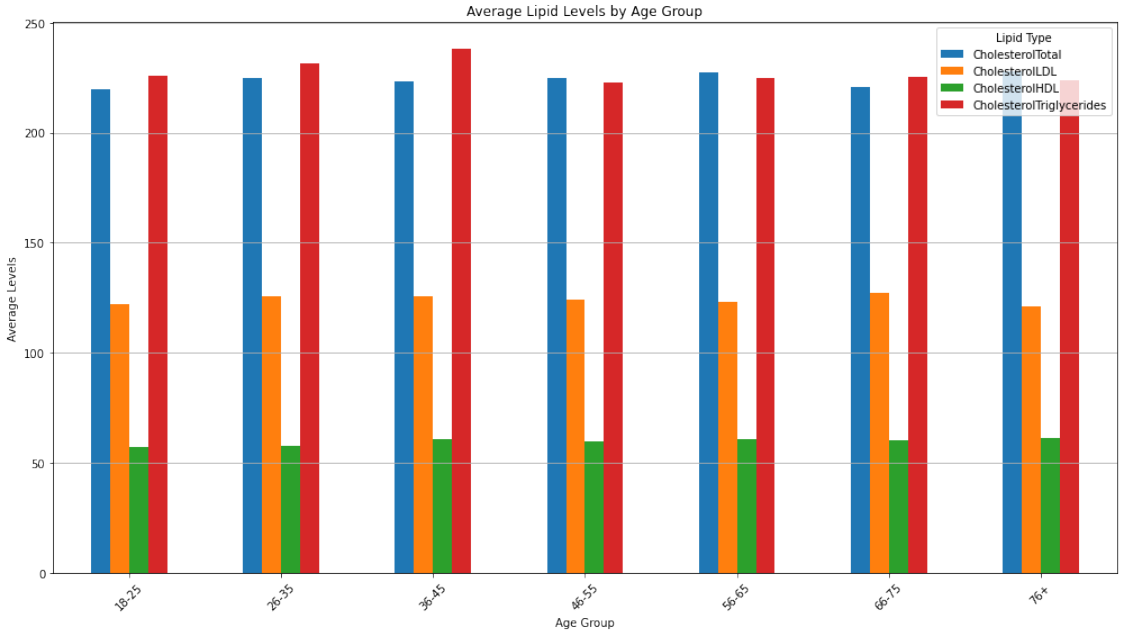
[58]:
```
# How do total cholesterol, LDL, HDL, and triglycerides compare among different␣
 ↪age groups?
# Creating age groups
bins = [18, 25, 35, 45, 55, 65, 75, 85]
labels = ['18-25', '26-35', '36-45', '46-55', '56-65', '66-75', '76+']
data['AgeGroup'] = pd.cut(data['Age'], bins=bins, labels=labels, right=False)
```

[64]:
```
# Calculating average lipid levels by age group
avg_lipidlevels = data.groupby('AgeGroup')[['CholesterolTotal',␣
 ↪'CholesterolLDL', 'CholesterolHDL', 'CholesterolTriglycerides']].mean().
 ↪reset_index()
```

[65]:
```
#How do total cholesterol, LDL, HDL, and triglycerides compare among different␣
 ↪age groups?
# Bar Plot for Average Lipid Levels
plt.figure(figsize=(14, 8))
avg_lipidlevels.plot(x='AgeGroup', kind='bar', figsize=(14, 8), legend=True)
plt.title('Average Lipid Levels by Age Group')
plt.xlabel('Age Group')
```

```
plt.ylabel('Average Levels')
plt.xticks(rotation=45)
plt.grid(axis='y')
plt.legend(title='Lipid Type')
plt.tight_layout()
plt.show()
```

```
<Figure size 1008x576 with 0 Axes>
```



[ ]: