zmnka12xr

April 12, 2025

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[3]:
```
data = pd.read_csv('Huntington_Disease_Dataset.csv')
data.head(5)
```

[3]:

	Patient_ID	Age	Sex	Family_History	HTT_CAG_Repeat_Length	Motor_Symptoms	Cognitive_Decline	Chorea_Score	Brain_Volume_Loss	Functional_Capacity	…	HTT_Gene_Expression_Level	Protein_Aggregation_Level	Random_Protein_Sequence
0	b2a49170-8561-4665-9371-2240b55dd87a	31	Male	No	67	Moderate	Severe	8.80	3.20	94	…	1.67	0.58	
1	f5fae45d-8718-41c4-a957-6928f79f3c8e	33	Female	Yes	38	Severe	Moderate	3.24	5.98	50	…	0.18	0.30	
2	66ab0567-050b-4d56-9ec4-b676309899a6	69	Male	Yes	37	Severe	Moderate	1.01	2.82	69	…	0.90	1.04	
3	996a48e4-e841-418f-a539-5a7a86cd815d	66	Male	Yes	50	Mild	Severe	3.21	6.77	76	…	1.16	1.87	
4	d45c7ca8-7125-4aaa-8018-5bbc60d35a1f	43	Female	Yes	48	Moderate	Mild	2.31	7.53	70	…	1.85	2.94	

```
0  DAHKIRSPMRVGPHYYAQCDNNDTGSDKEHWLKTEAAPMTMDRTVE…
1  PANGFWYHNCLRFWNIPPYVMEGFPLADITEVHKWRVSGFMCWETQ…
2  NWHEGHGASTWKATMVAWCLMVQHAVTWKEGNTRCREMSCMNFTQL…
3  KCVQYIQATQMLVQSWGQRNPIMQSSEPDRAHDYESGTPKTYTYML…
4  DQPGNMTRQNKNHCMWRAKRPTKHPGHKPGEIDKEKSEQNDADSSA…

                              Random_Gene_Sequence    Disease_Stage  \
0  GCCAGCAGCGCCCGAGCGTATGAGGTATATGGATTGGACATTGGGC…           Middle
1  AGTTTTCAGTGAGACTCTTCCCCAAAAGCCTCCACTACGACAGTGT…  Pre-Symptomatic
2  TATACCACCAGTGGGAAGAGTAACGATTTTGGAGCGCCCCGAGTCC…            Early
3  GCGCGACCGACCAAAGGACCCATGGTGGTGATCTGTCATTGGATTC…  Pre-Symptomatic
4  GGGACCGCGGTTCTAGAAGAGAGGTTCTCTGACCGCCGAAGGATTC…             Late

              Gene/Factor Chromosome_Location  \
0                     HTT              4p16.3
1                     HTT              4p16.3
2                    MSH3              5q14.1
3                    MSH3              5q14.1
4  HTT (Somatic Expansion)              4p16.3

                           Function                Effect  \
0  CAG Trinucleotide Repeat Expansion     Neurodegeneration
1  CAG Trinucleotide Repeat Expansion     Neurodegeneration
2                   Mismatch Repair  CAG Repeat Expansion
3                   Mismatch Repair  CAG Repeat Expansion
4             CAG Repeat Instability  Faster Disease Onset

               Category
0         Primary Cause
1         Primary Cause
2  Trans-acting Modifier
3  Trans-acting Modifier
4    Cis-acting Modifier

[5 rows x 21 columns]
```

[4]:
```
data.describe()
```

[4]:
```
                Age  HTT_CAG_Repeat_Length  Chorea_Score  Brain_Volume_Loss  \
count  48536.000000           48536.000000  48536.000000       48536.000000
mean      55.070566              57.516606      4.994399           5.261030
std       14.762154              13.264153      2.897327           1.872995
min       30.000000              35.000000      0.000000           2.000000
25%       42.000000              46.000000      2.480000           3.640000
50%       55.000000              58.000000      4.980000           5.280000
75%       68.000000              69.000000      7.500000           6.880000
max       80.000000              80.000000     10.000000           8.500000
```

```
       Functional_Capacity  HTT_Gene_Expression_Level  \
count         48536.000000                48536.000000
mean             50.354829                    1.301060
std              29.189697                    0.691657
min               0.000000                    0.100000
25%              25.000000                    0.700000
50%              50.000000                    1.300000
75%              76.000000                    1.900000
max             100.000000                    2.500000

       Protein_Aggregation_Level
count                48536.000000
mean                     2.546519
std                      1.414745
min                      0.100000
25%                      1.330000
50%                      2.540000
75%                      3.770000
max                      5.000000
```

[5]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48536 entries, 0 to 48535
Data columns (total 21 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Patient_ID                 48536 non-null  object
 1   Age                        48536 non-null  int64
 2   Sex                        48536 non-null  object
 3   Family_History             48536 non-null  object
 4   HTT_CAG_Repeat_Length      48536 non-null  int64
 5   Motor_Symptoms             48536 non-null  object
 6   Cognitive_Decline          36417 non-null  object
 7   Chorea_Score               48536 non-null  float64
 8   Brain_Volume_Loss          48536 non-null  float64
 9   Functional_Capacity        48536 non-null  int64
 10  Gene_Mutation_Type         48536 non-null  object
 11  HTT_Gene_Expression_Level  48536 non-null  float64
 12  Protein_Aggregation_Level  48536 non-null  float64
 13  Random_Protein_Sequence    48536 non-null  object
 14  Random_Gene_Sequence       48536 non-null  object
 15  Disease_Stage              48536 non-null  object
 16  Gene/Factor                48536 non-null  object
 17  Chromosome_Location        48536 non-null  object
 18  Function                   48536 non-null  object
```

```
 19  Effect                     48536 non-null  object
 20  Category                   48536 non-null  object
dtypes: float64(4), int64(3), object(14)
memory usage: 7.8+ MB
```

[6]:
```
data.shape
```

[6]:
```
(48536, 21)
```

[7]:
```
data.tail(5)
```

[7]:

	Patient_ID	Age	Sex	Family_History	HTT_CAG_Repeat_Length	Motor_Symptoms	Cognitive_Decline	Chorea_Score	Brain_Volume_Loss	Functional_Capacity	…	HTT_Gene_Expression_Level	Protein_Aggregation_Level	Random_Protein_Sequence	Random_Gene_Sequence	Disease_Stage
48531	27f82b73-b9fe-49d6-b6b9-82e667054725	78	Female	Yes	38	Moderate	Severe	0.09	4.71	89	…	0.49	1.24	WGYQYSFEMVIQWMYHSYMPQGGLGCHNENHYCFWQDHKVATAVTC…		
48532	ff5af7cc-8132-4791-8e8b-aac71f2bade1	61	Female	Yes	76	Mild	NaN	8.80	6.92	78	…	2.44	3.34	TNVLFHILGLVENNMGQKFHQKAFPWCDTLMTSNWNGCAEGGNGEQ…		
48533	f594aab8-7acf-43cd-95d8-c197e202c4ea	66	Male	No	35	Severe	Severe	2.73	6.61	24	…	0.51	0.27	NRWNVDINSRMMHFYYFHTSPRSVTPWTPKQEREVYKQMFQKYAQS…		
48534	32052e20-c2ae-4314-95ce-5daf720b26ca	37	Female	Yes	80	Mild	NaN	9.81	2.95	19	…	1.16	3.22	AKDHQGVPWLPRITQKQTVQQQPIKWMHWVPENARGTIIREFFAYA…		
48535	b063b8fb-9e12-40c2-b6bb-da675033dfd2	37	Female	Yes	48	Severe	NaN	0.26	5.25	50	…	1.70	4.59	WTRLKMNYRQFKRERKYEGQEAIAPSEKGKDNAFPTMEIGIWCNES…		

```
48531  CAATGACGCACCCTCGTAAGGGCTAGGCAGCGCCAGCCCACTTAAA…         Early
48532  ACAAGAACAGTGGAACGGTTGAGTCACGGATGAGGCCTAGCGAGGT…        Middle
48533  GGTATATCGTTAGGGCCGTCGGTGAACACTTCTAATCTAGGCTTAA…        Middle
48534  GGTTTTCGTTCTCGTAGTATCAACAGAGGGGAATTTTATTACTGGC…          Late
48535  ACCGGCGGAGACCGCCGATGTAGTGGTCCTCACTGACCATCTCGCA…         Early

                  Gene/Factor Chromosome_Location                Function  \
48531                    MSH3              5q14.1         Mismatch Repair
48532  HTT (Somatic Expansion)              4p16.3  CAG Repeat Instability
48533                    MSH3              5q14.1         Mismatch Repair
48534  HTT (Somatic Expansion)              4p16.3  CAG Repeat Instability
48535  HTT (Somatic Expansion)              4p16.3  CAG Repeat Instability

                     Effect               Category
48531  CAG Repeat Expansion  Trans-acting Modifier
48532  Faster Disease Onset    Cis-acting Modifier
48533  CAG Repeat Expansion  Trans-acting Modifier
48534  Faster Disease Onset    Cis-acting Modifier
48535  Faster Disease Onset    Cis-acting Modifier

[5 rows x 21 columns]
```

[8]:
```
data.isnull().sum()
```

[8]:
```
Patient_ID                     0
Age                            0
Sex                            0
Family_History                 0
HTT_CAG_Repeat_Length          0
Motor_Symptoms                 0
Cognitive_Decline          12119
Chorea_Score                   0
Brain_Volume_Loss              0
Functional_Capacity            0
Gene_Mutation_Type             0
HTT_Gene_Expression_Level      0
Protein_Aggregation_Level      0
Random_Protein_Sequence        0
Random_Gene_Sequence           0
Disease_Stage                  0
Gene/Factor                    0
Chromosome_Location            0
Function                       0
Effect                         0
Category                       0
dtype: int64
```

[9]:
```
data.duplicated().sum()
```

[9]:
```
np.int64(0)
```

[10]:
```
data_num = data.select_dtypes(include=np.number)
data_num.head(5)
```

[10]:

	Age	HTT_CAG_Repeat_Length	Chorea_Score	Brain_Volume_Loss	Functional_Capacity	HTT_Gene_Expression_Level	Protein_Aggregation_Level
0	31	67	8.80	3.20	94	1.67	0.58
1	33	38	3.24	5.98	50	0.18	0.30
2	69	37	1.01	2.82	69	0.90	1.04
3	66	50	3.21	6.77	76	1.16	1.87
4	43	48	2.31	7.53	70	1.85	2.94

[18]:
```
data_object = data.select_dtypes(include='object')
data_object.head(5)
```

[18]:

	Patient_ID	Sex	Family_History	Motor_Symptoms	Gene_Mutation_Type	Random_Protein_Sequence	Random_Gene_Sequence	Disease_Stage	Gene/Factor	Chromosome_Location
0	b2a49170-8561-4665-9371-2240b55dd87a	Male	No	Moderate	Deletion	DAHKIRSPMRVGPHYYAQCDNNDTGSDKEHWLKTEAAPMTMDRTVE…	GCCAGCAGCGCCCGAGCGTATGAGGTATATGGATTGGACATTGGGC…	Middle	HTT	4p16.3
1	f5fae45d-8718-41c4-a957-6928f79f3c8e	Female	Yes	Severe	Point Mutation	PANGFWYHNCLRFWNIPPYVMEGFPLADITEVHKWRVSGFMCWETQ…	AGTTTTCAGTGAGACTCTTCCCCAAAAGCCTCCACTACGACAGTGT…	Pre-Symptomatic		
2	66ab0567-050b-4d56-9ec4-b676309899a6	Male	Yes	Severe	Duplication	NWHEGHGASTWKATMVAWCLMVQHAVTWKEGNTRCREMSCMNFTQL…	TATACCACCAGTGGGAAGAGTAACGATTTTGGAGCGCCCCGAGTCC…	Early		
3	996a48e4-e841-418f-a539-5a7a86cd815d	Male	Yes	Mild	Deletion	KCVQYIQATQMLVQSWGQRNPIMQSSEPDRAHDYESGTPKTYTYML…	GCGCGACCGACCAAAGGACCCATGGTGGTGATCTGTCATTGGATTC…	Pre-Symptomatic		
4	d45c7ca8-7125-4aaa-8018-5bbc60d35a1f	Female	Yes	Moderate	Insertion	DQPGNMTRQNKNHCMWRAKRPTKHPGHKPGEIDKEKSEQNDADSSA…	GGGACCGCGGTTCTAGAAGAGAGGTTCTCTGACCGCCGAAGGATTC…	Late		

```
1                      HTT              4p16.3
2                     MSH3              5q14.1
3                     MSH3              5q14.1
4  HTT (Somatic Expansion)              4p16.3

                            Function                Effect  \
0  CAG Trinucleotide Repeat Expansion     Neurodegeneration
1  CAG Trinucleotide Repeat Expansion     Neurodegeneration
2                     Mismatch Repair  CAG Repeat Expansion
3                     Mismatch Repair  CAG Repeat Expansion
4              CAG Repeat Instability  Faster Disease Onset

                Category
0          Primary Cause
1          Primary Cause
2  Trans-acting Modifier
3  Trans-acting Modifier
4    Cis-acting Modifier
```

[11]:
```
data_num.columns
```

[11]:
```
Index(['Age', 'HTT_CAG_Repeat_Length', 'Chorea_Score', 'Brain_Volume_Loss',
       'Functional_Capacity', 'HTT_Gene_Expression_Level',
       'Protein_Aggregation_Level'],
      dtype='object')
```

[12]:
```
data.drop('Cognitive_Decline', axis=1, inplace=True)
# column was already dropped
```

[99]:
```
# save to csv
data.to_csv('Huntington_Disease_Dataset_Cleaned.csv', index=False)
```

[13]:
```
data.columns
```

[13]:
```
Index(['Patient_ID', 'Age', 'Sex', 'Family_History', 'HTT_CAG_Repeat_Length',
       'Motor_Symptoms', 'Chorea_Score', 'Brain_Volume_Loss',
       'Functional_Capacity', 'Gene_Mutation_Type',
       'HTT_Gene_Expression_Level', 'Protein_Aggregation_Level',
       'Random_Protein_Sequence', 'Random_Gene_Sequence', 'Disease_Stage',
       'Gene/Factor', 'Chromosome_Location', 'Function', 'Effect', 'Category'],
      dtype='object')
```

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

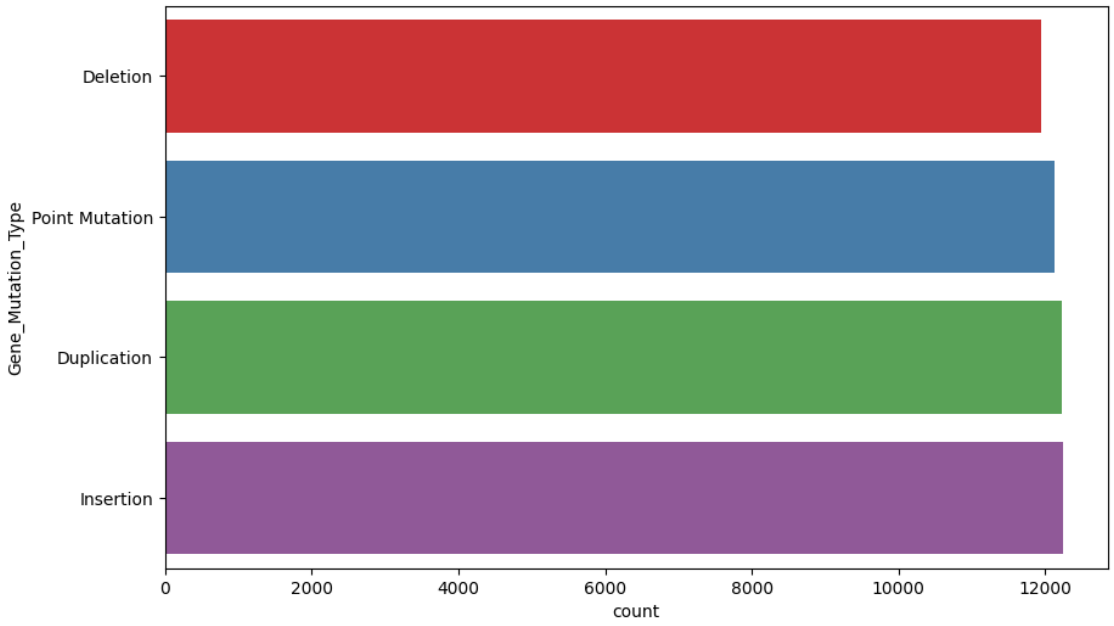
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48536 entries, 0 to 48535
Data columns (total 20 columns):
```

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Patient_ID                 48536 non-null  object
 1   Age                        48536 non-null  int64
 2   Sex                        48536 non-null  object
 3   Family_History             48536 non-null  object
 4   HTT_CAG_Repeat_Length      48536 non-null  int64
 5   Motor_Symptoms             48536 non-null  object
 6   Chorea_Score               48536 non-null  float64
 7   Brain_Volume_Loss          48536 non-null  float64
 8   Functional_Capacity        48536 non-null  int64
 9   Gene_Mutation_Type         48536 non-null  object
 10  HTT_Gene_Expression_Level  48536 non-null  float64
 11  Protein_Aggregation_Level  48536 non-null  float64
 12  Random_Protein_Sequence    48536 non-null  object
 13  Random_Gene_Sequence       48536 non-null  object
 14  Disease_Stage              48536 non-null  object
 15  Gene/Factor                48536 non-null  object
 16  Chromosome_Location        48536 non-null  object
 17  Function                   48536 non-null  object
 18  Effect                     48536 non-null  object
 19  Category                   48536 non-null  object
dtypes: float64(4), int64(3), object(13)
memory usage: 7.4+ MB
```

0.0.1 VISUALIZATION

[19]:
```
plt.figure(figsize=(10,6))
sns.countplot(data['Gene_Mutation_Type'], palette='Set1')
plt.show()
```

```
<ipython-input-19-ef8926482d6a>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data['Gene_Mutation_Type'], palette='Set1')
```



[21]:
```
data_object.columns
```

[21]:
```
Index(['Patient_ID', 'Sex', 'Family_History', 'Motor_Symptoms',
       'Gene_Mutation_Type', 'Random_Protein_Sequence', 'Random_Gene_Sequence',
       'Disease_Stage', 'Gene/Factor', 'Chromosome_Location', 'Function',
       'Effect', 'Category'],
      dtype='object')
```

[23]:
```
data['Gene/Factor'].value_counts()
```

[23]:
```
Gene/Factor
HTT                          12229
MLH1                         12223
MSH3                         12056
HTT (Somatic Expansion)      12028
Name: count, dtype: int64
```

[24]:
```
data.describe()
```

[24]:

	Age	HTT_CAG_Repeat_Length	Chorea_Score	Brain_Volume_Loss	\
count	48536.000000	48536.000000	48536.000000	48536.000000	
mean	55.070566	57.516606	4.994399	5.261030	
std	14.762154	13.264153	2.897327	1.872995	
min	30.000000	35.000000	0.000000	2.000000	
25%	42.000000	46.000000	2.480000	3.640000	
50%	55.000000	58.000000	4.980000	5.280000	

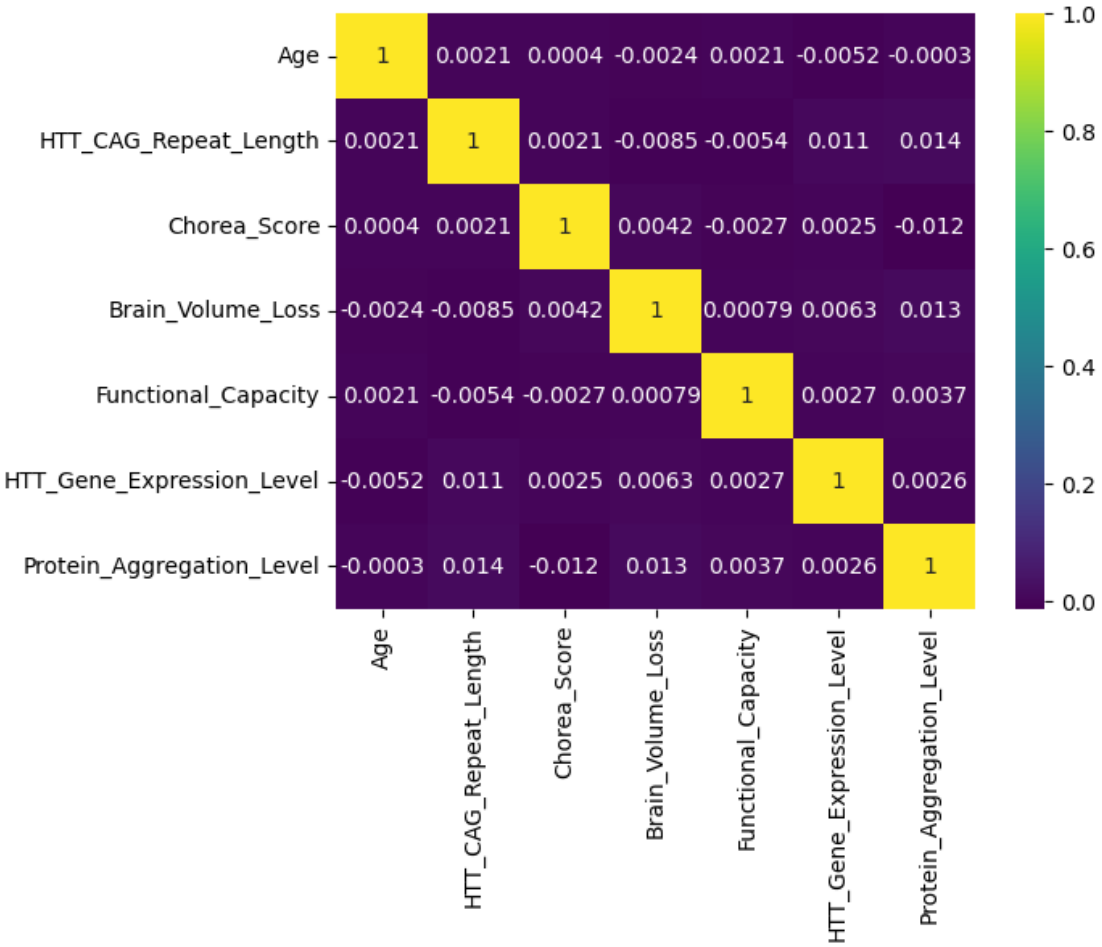
```
75%       68.000000              69.000000       7.500000           6.880000
max       80.000000              80.000000      10.000000           8.500000

       Functional_Capacity  HTT_Gene_Expression_Level  \
count         48536.000000                48536.000000
mean             50.354829                    1.301060
std              29.189697                    0.691657
min               0.000000                    0.100000
25%              25.000000                    0.700000
50%              50.000000                    1.300000
75%              76.000000                    1.900000
max             100.000000                    2.500000

       Protein_Aggregation_Level
count                48536.000000
mean                     2.546519
std                      1.414745
min                      0.100000
25%                      1.330000
50%                      2.540000
75%                      3.770000
max                      5.000000
```

[25]:
```
corr_matrix = data_num.corr()
corr_matrix
```

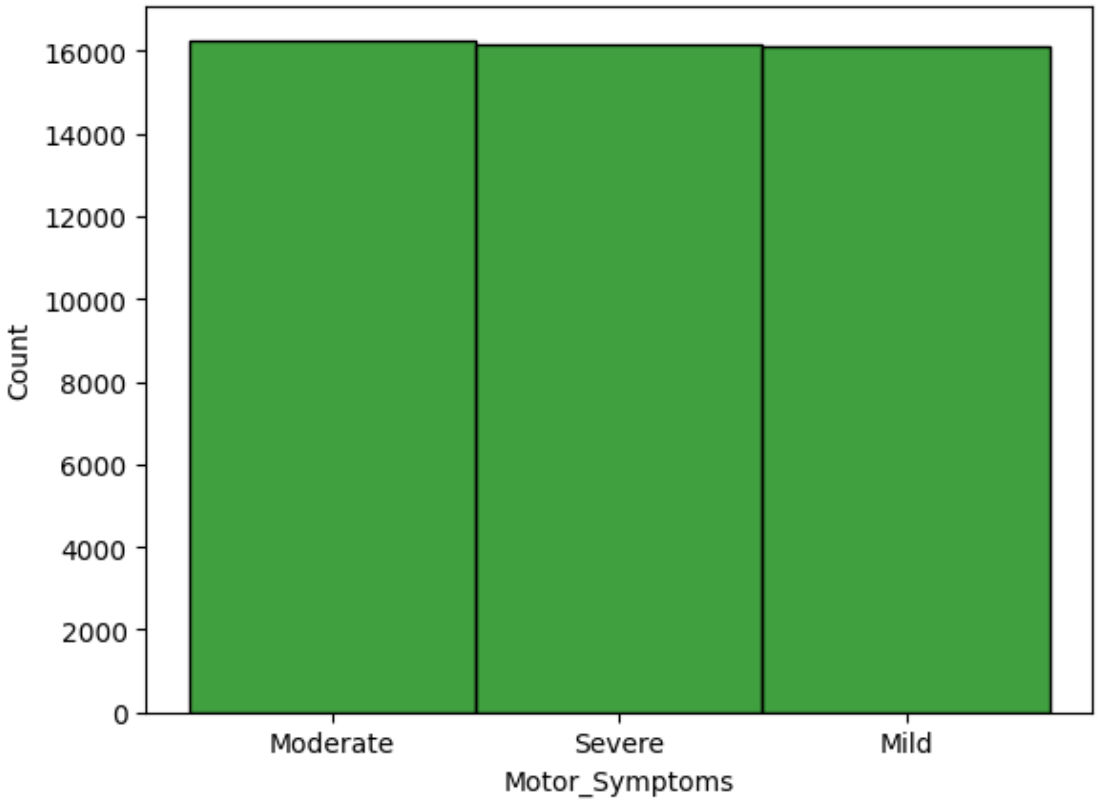
[25]:
```
                                Age  HTT_CAG_Repeat_Length  Chorea_Score  \
Age                        1.000000               0.002051      0.000405
HTT_CAG_Repeat_Length      0.002051               1.000000      0.002078
Chorea_Score               0.000405               0.002078      1.000000
Brain_Volume_Loss         -0.002439              -0.008470      0.004227
Functional_Capacity        0.002096              -0.005382     -0.002651
HTT_Gene_Expression_Level -0.005172               0.011170      0.002485
Protein_Aggregation_Level -0.000304               0.014360     -0.012492

                           Brain_Volume_Loss  Functional_Capacity  \
Age                                -0.002439             0.002096
HTT_CAG_Repeat_Length              -0.008470            -0.005382
Chorea_Score                        0.004227            -0.002651
Brain_Volume_Loss                   1.000000             0.000790
Functional_Capacity                 0.000790             1.000000
HTT_Gene_Expression_Level           0.006296             0.002689
Protein_Aggregation_Level           0.012553             0.003715

                           HTT_Gene_Expression_Level  \
Age                                        -0.005172
HTT_CAG_Repeat_Length                       0.011170
```

```
Chorea_Score                          0.002485
Brain_Volume_Loss                     0.006296
Functional_Capacity                   0.002689
HTT_Gene_Expression_Level             1.000000
Protein_Aggregation_Level             0.002638

                          Protein_Aggregation_Level
Age                                       -0.000304
HTT_CAG_Repeat_Length                      0.014360
Chorea_Score                              -0.012492
Brain_Volume_Loss                          0.012553
Functional_Capacity                        0.003715
HTT_Gene_Expression_Level                  0.002638
Protein_Aggregation_Level                  1.000000
```

[27]:
```
sns.heatmap(corr_matrix,cmap='viridis',annot=True)
plt.show()
```

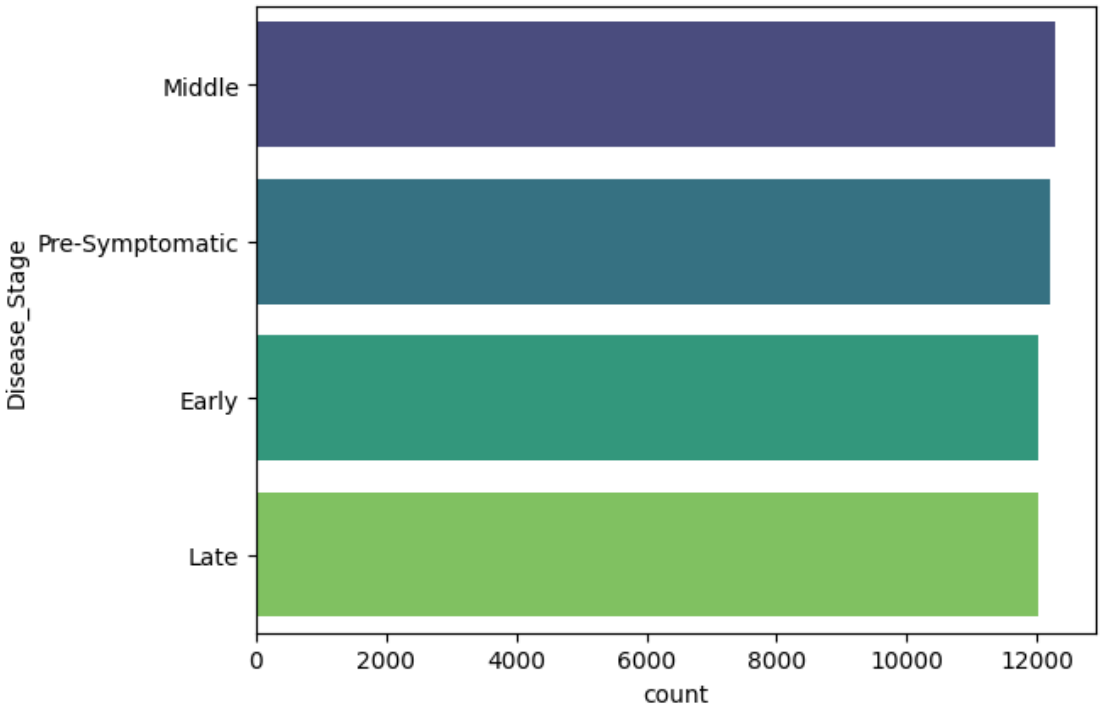


[108]:
```
# motor symptoms
sns.histplot(data['Motor_Symptoms']
             ,color='green')
plt.show()
```



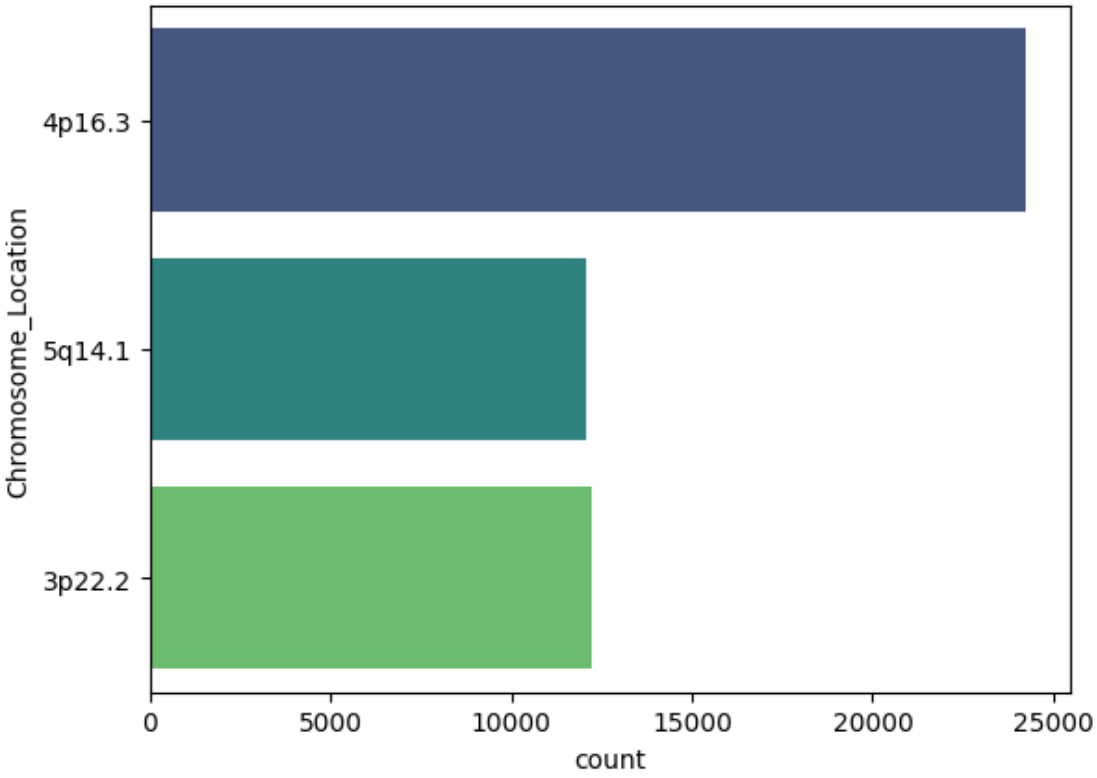
[34]:
```
# disease stage
sns.countplot(data['Disease_Stage'], palette ='viridis')
plt.show()
```

```
<ipython-input-34-c6a0133d7b38>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data['Disease_Stage'], palette ='viridis')
```



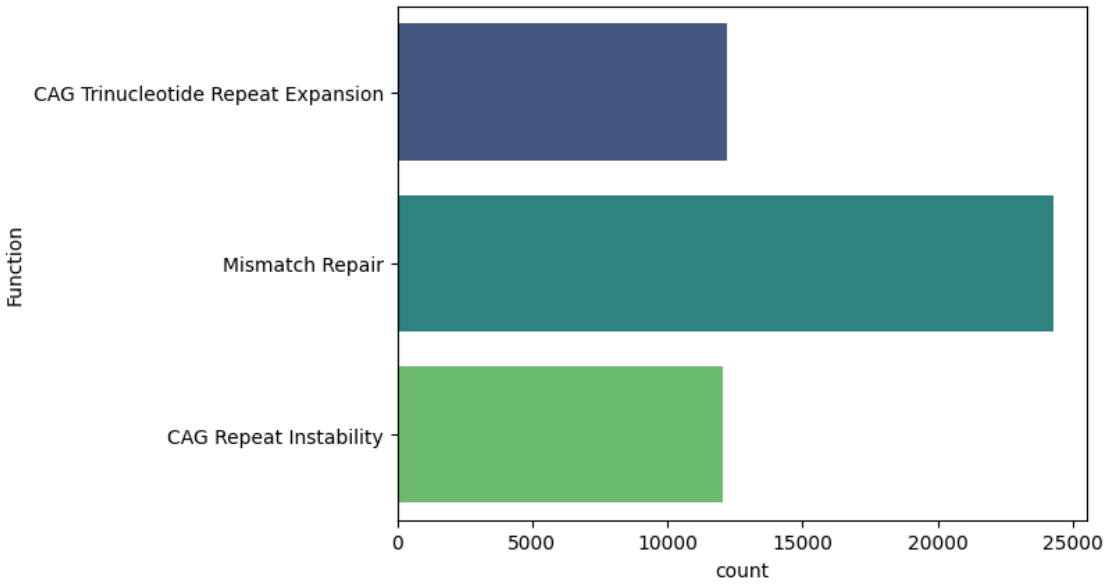
[35]:
```
# Chromosome Location
sns.countplot(data['Chromosome_Location'], palette='viridis')
plt.show()
```

```
<ipython-input-35-9ca3a8cb2123>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data['Chromosome_Location'], palette='viridis')
```



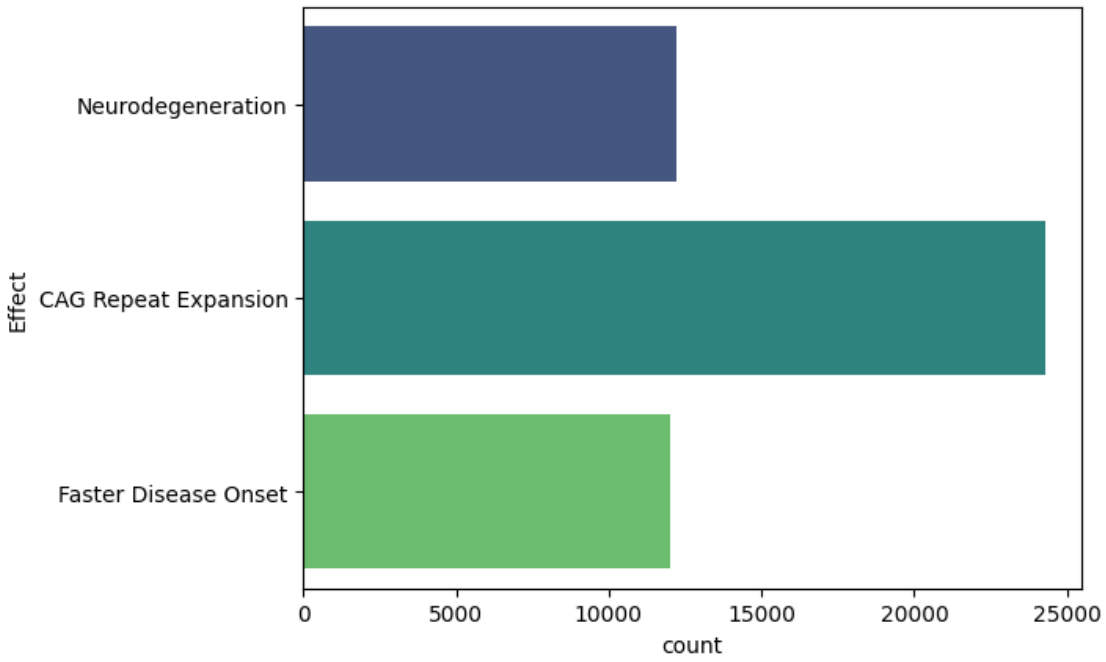
[36]:
```
# Function
sns.countplot(data['Function'], palette='viridis')
plt.show()
```

```
<ipython-input-36-638455c276e0>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data['Function'], palette='viridis')
```



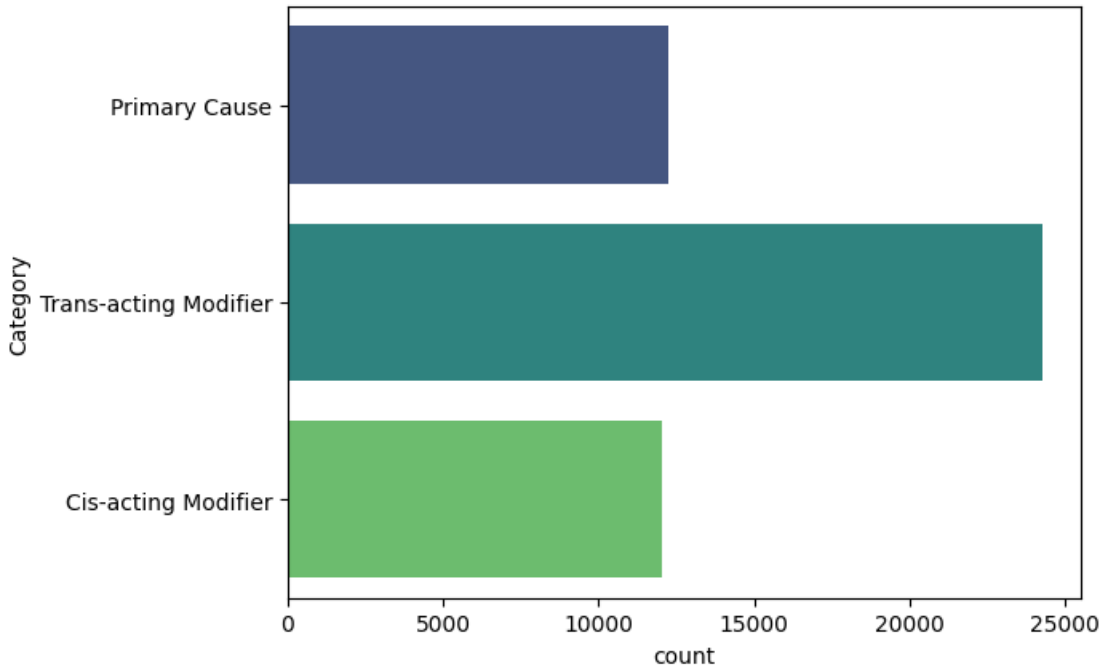
[37]:
```
# Effect
sns.countplot(data['Effect'], palette='viridis')
plt.show()
```

```
<ipython-input-37-9e765ab10577>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data['Effect'], palette='viridis')
```



[38]:
```
# Category
sns.countplot(data['Category'], palette='viridis')
plt.show()
```

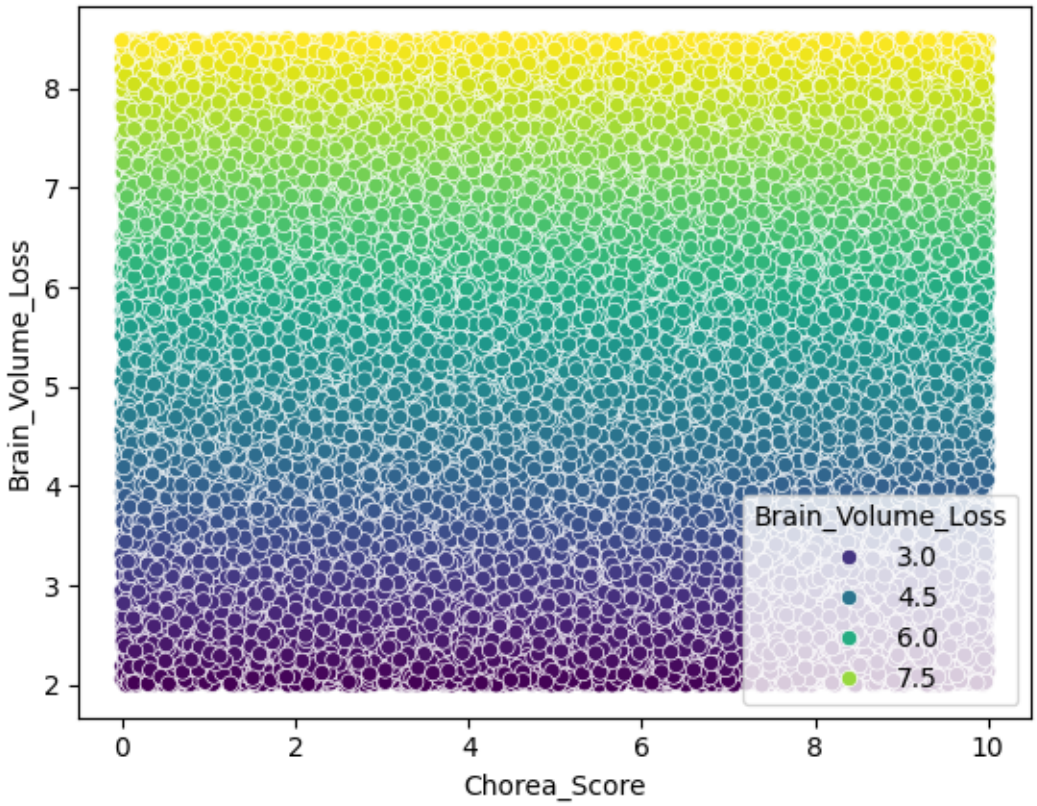
```
<ipython-input-38-c4ce8c0a8244>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data['Category'], palette='viridis')
```



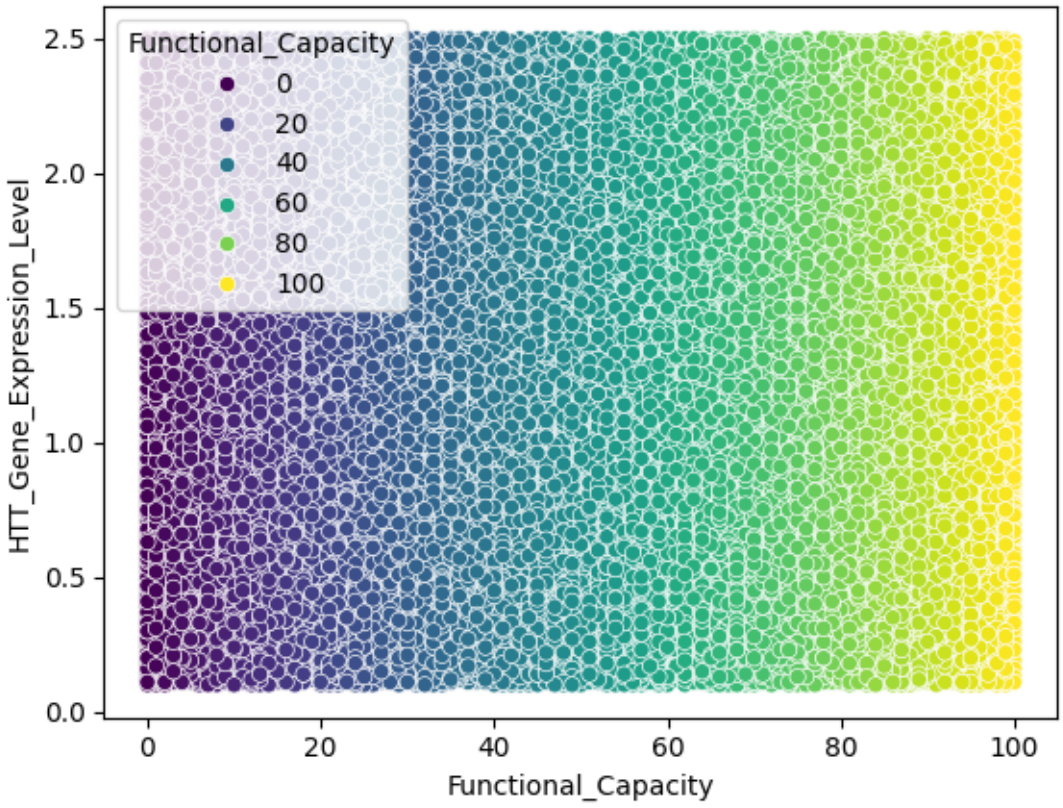
[28]:
```
data_num.columns
```

[28]:
```
Index(['Age', 'HTT_CAG_Repeat_Length', 'Chorea_Score', 'Brain_Volume_Loss',
       'Functional_Capacity', 'HTT_Gene_Expression_Level',
       'Protein_Aggregation_Level'],
      dtype='object')
```

[44]:
```
sns.scatterplot(data=data, x='Chorea_Score', y='Brain_Volume_Loss',␣
↪hue='Brain_Volume_Loss', palette='viridis')
plt.show()
```



[46]:
```
# functional capacity vs gene expression level
sns.scatterplot(x='Functional_Capacity', y='HTT_Gene_Expression_Level',␣
 ↪data=data, hue='Functional_Capacity', palette='viridis')
plt.show()
```



[62]:
```
data_num.columns
```

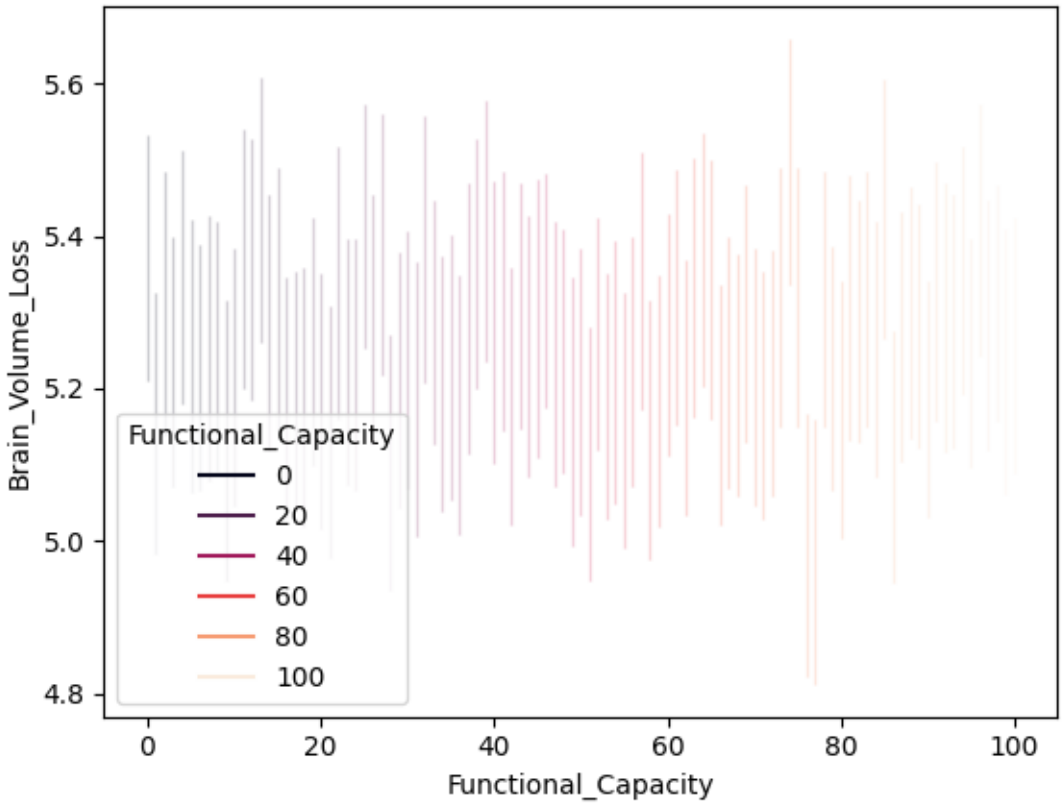
[62]:
```
Index(['Age', 'HTT_CAG_Repeat_Length', 'Chorea_Score', 'Brain_Volume_Loss',
       'Functional_Capacity', 'HTT_Gene_Expression_Level',
       'Protein_Aggregation_Level'],
      dtype='object')
```

[64]:
```
data['Functional_Capacity'].value_counts()
data['Functional_Capacity'].unique()
```

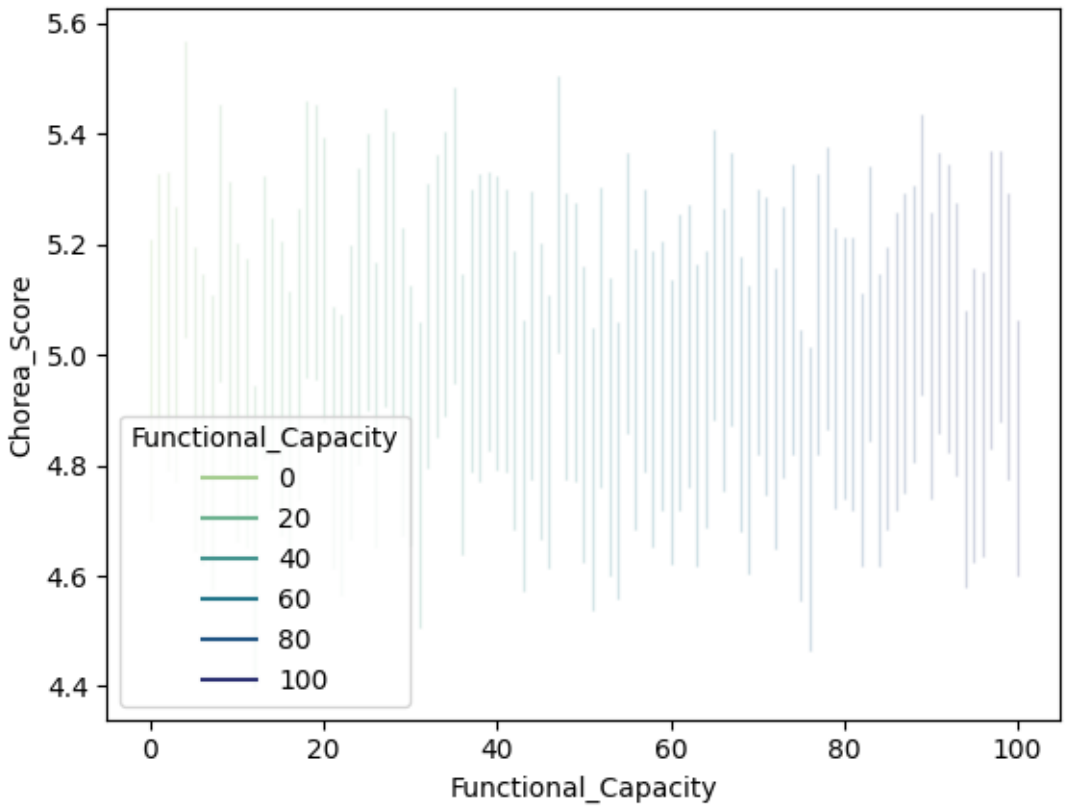
[64]:
```
array([ 94,  50,  69,  76,  70,  78,  87,   3,  20,  39,  75,  86,  82,
        14,  55,  95,  57,  67,  40,  27,   2,  34,  89,   9,  90,   6,
        26,  28,   7,  43,  19,  31,  92,  84,   1,  13,  44,  68,  41,
        25,  38,  66,  72,  17, 100,  52,  83,  91,  88,  64,  62,  18,
        49,  97,  24,  77,  45,  35,  58,  36,  42,  98,  80,  21,  60,
        47,  30,  16,  51,  93,  33,  23,  85,   4,  79,  71,  53,   8,
        81,  56,  29,  11,  22,  59,  65,  46,  10,  15,  74,  61,   5,
        37,  54,  63,   0,  99,  32,  48,  73,  12,  96])
```

[69]:
```
sns.lineplot(x='Functional_Capacity', y='Brain_Volume_Loss', data=data,␣
↪hue='Functional_Capacity', palette='rocket')
```

```
plt.show()
```

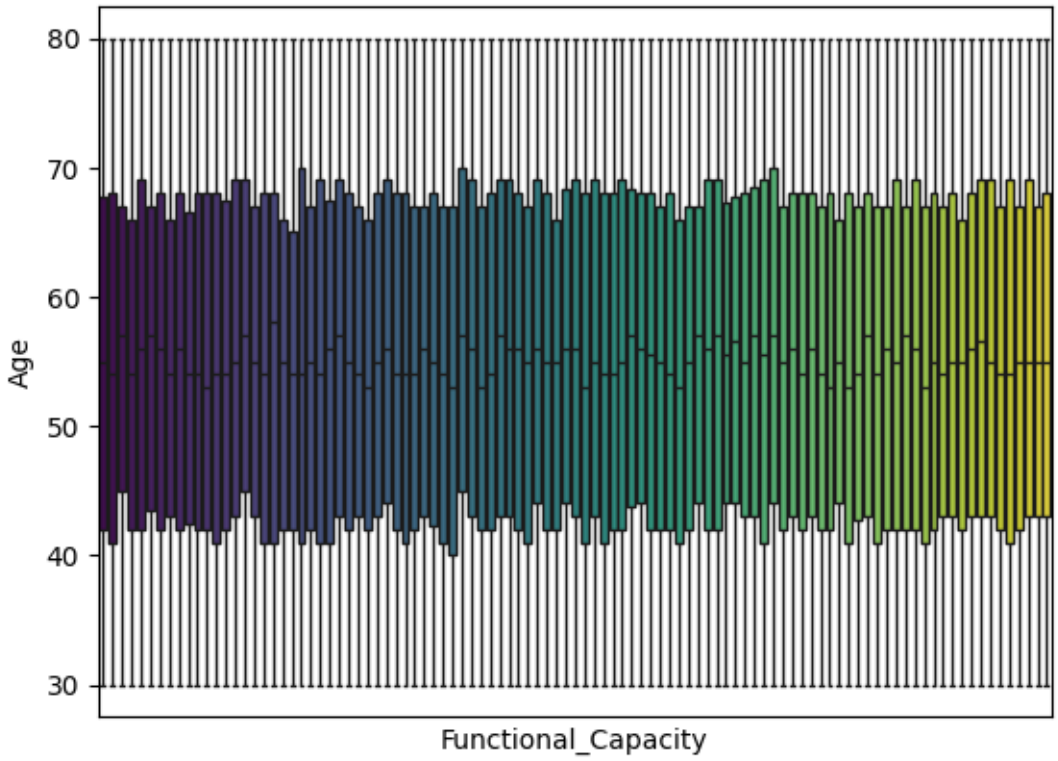


[71]:
```
sns.lineplot(x='Functional_Capacity', y='Chorea_Score', data=data,␣
↪hue='Functional_Capacity', palette='crest')
plt.show()
```

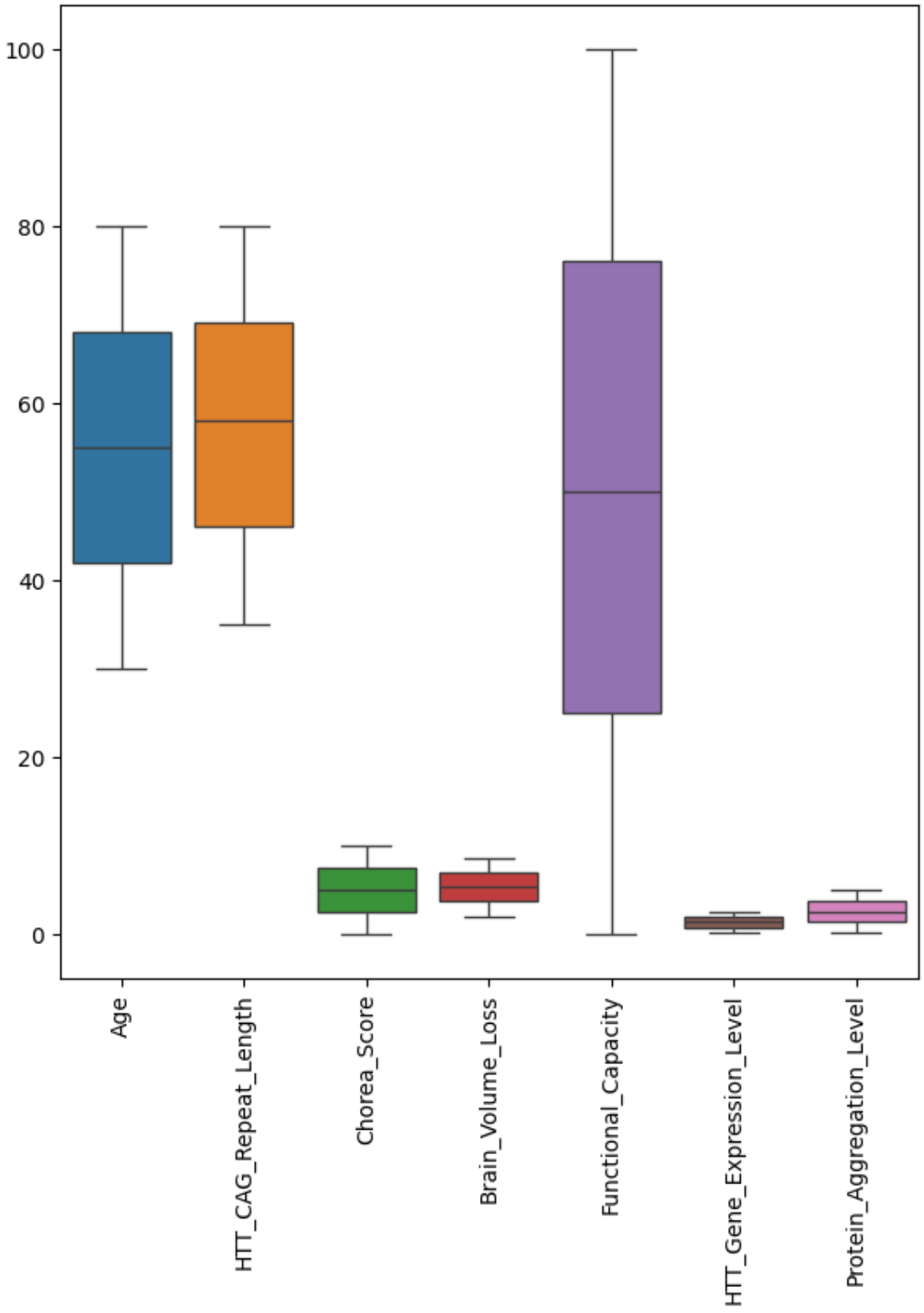


[74]:
```
sns.boxplot(x='Functional_Capacity', y='Age', data=data, palette='viridis')
plt.xticks([])
plt.show()
```

```
<ipython-input-74-1314b5986a95>:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  sns.boxplot(x='Functional_Capacity', y='Age', data=data, palette='viridis')
```



[82]:
```
# outlier detection
plt.figure(figsize=(7,8))
sns.boxplot(data=data_num)
plt.xticks(rotation=90)
plt.show()
```


100
80
60
40
20
0
Age
HTT_CAG_Repeat_Length
Chorea_Score
Brain_Volume_Loss
Functional_Capacity
HTT_Gene_Expression_Level
Protein_Aggregation_Level

[93]:
```
from scipy import stats
z_score = np.abs(stats.zscore(data_num.head(5)))
print(z_score)
```

```
        Age  HTT_CAG_Repeat_Length  Chorea_Score  Brain_Volume_Loss  \
0  1.077777               1.755051      1.905285           1.081261
1  0.953894               0.923711      0.177567           0.377917
2  1.275988               1.016082      1.012955           1.280717
3  1.090165               0.184742      0.188805           0.792575
4  0.334482               0.000000      0.525958           1.191487

   Functional_Capacity  HTT_Gene_Expression_Level  Protein_Aggregation_Level
0             1.572295                   0.872128                   0.799487
1             1.543965                   1.636503                   1.091727
2             0.198307                   0.424279                   0.319377
3             0.297461                   0.013469                   0.546907
4             0.127483                   1.175184                   1.663684
```

[95]:
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data_num)
scaled_data
```

[95]:
```
array([[-1.63057598,  0.71497152,  1.3135005 , …,  1.49524059,
         0.53342011, -1.39003127],
       [-1.49509301, -1.47139494, -0.60552968, …, -0.01215611,
        -1.62085021, -1.58794884],
       [ 0.94360058, -1.54678689, -1.37521265, …,  0.63876519,
        -0.57986052, -1.06488098],
       …,
       [ 0.74037612, -1.69757078, -0.78155583, …, -0.90289052,
        -1.14372993, -1.60915429],
       [-1.22412705,  1.69506683,  1.66210131, …, -1.0741856 ,
        -0.20394758,  0.47604866],
       [-1.22412705, -0.71747547, -1.63407464, …, -0.01215611,
         0.57679468,  1.44443105]])
```

[104]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

X = data[['Brain_Volume_Loss']]
y = data['Chorea_Score']

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.2,␣
 ↪random_state=42)
```

```
model = LinearRegression()
model.fit(X_train,y_train)
y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mean Squared Error: ", mse)
print("R-squared: ",r2)
```

```
Mean Squared Error:  8.37915178387766
R-squared:  -0.00011636900441902576
```

[ ]: