yhg5xxu6d

April 23, 2025

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
df = pd.read_csv('train.csv.zip')
df.head()
```

[2]:

	Patient Id	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Patient First Name	Family Name	Father's name	…	Birth defects	White Blood cell count (thousand per microliter)	Blood test result
0	PID0x6418	2.0	Yes	No	Yes	No	4.760603	Richard	NaN	Larre	…	NaN	9.857562	NaN
1	PID0x25d5	4.0	Yes	Yes	No	No	4.910669	Mike	NaN	Brycen	…	Multiple	5.522560	normal
2	PID0x4a82	6.0	Yes	No	No	No	4.893297	Kimberly	NaN	Nashon	…	Singular	NaN	normal
3	PID0x4ac8	12.0	Yes	No	Yes	No	4.705280	Jeffery	Hoelscher	Aayaan	…	Singular	7.919321	inconclusive
4	PID0x1bf7	11.0	Yes	No	NaN	Yes	4.720703	Johanna	Stutzman	Suave	…	Multiple	4.098210	NaN

```
  Symptom 1 Symptom 2 Symptom 3 Symptom 4  Symptom 5  \
0       1.0       1.0       1.0       1.0        1.0
1       1.0       NaN       1.0       1.0        0.0
2       0.0       1.0       1.0       1.0        1.0
3       0.0       0.0       1.0       0.0        0.0
4       0.0       0.0       0.0       0.0        NaN

                                Genetic Disorder  \
0   Mitochondrial genetic inheritance disorders
1                                            NaN
2  Multifactorial genetic inheritance disorders
3   Mitochondrial genetic inheritance disorders
4  Multifactorial genetic inheritance disorders

                     Disorder Subclass
0  Leber's hereditary optic neuropathy
1                      Cystic fibrosis
2                             Diabetes
3                       Leigh syndrome
4                               Cancer

[5 rows x 45 columns]
```

[3]:
```
df.columns
```

[3]:
```
Index(['Patient Id', 'Patient Age', 'Genes in mother's side',
       'Inherited from father', 'Maternal gene', 'Paternal gene',
       'Blood cell count (mcL)', 'Patient First Name', 'Family Name',
       'Father's name', 'Mother's age', 'Father's age', 'Institute Name',
       'Location of Institute', 'Status', 'Respiratory Rate (breaths/min)',
       'Heart Rate (rates/min', 'Test 1', 'Test 2', 'Test 3', 'Test 4',
       'Test 5', 'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

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

[4]:
```
         Patient Age  Blood cell count (mcL)  Mother's age  Father's age  \
count  20656.000000            22083.000000  16047.000000  16097.000000
mean       6.974148                4.898871     34.526454     41.972852
std        4.319475                0.199663      9.852598     13.035501
min        0.000000                4.092727     18.000000     20.000000
25%        3.000000                4.763109     26.000000     31.000000
50%        7.000000                4.899399     35.000000     42.000000
75%       11.000000                5.033830     43.000000     53.000000
max       14.000000                5.609829     51.000000     64.000000

         Test 1   Test 2   Test 3   Test 4   Test 5  No. of previous abortion  \
count  19956.0  19931.0  19936.0  19943.0  19913.0              19921.000000
mean       0.0      0.0      0.0      1.0      0.0                  2.003062
std        0.0      0.0      0.0      0.0      0.0                  1.411919
min        0.0      0.0      0.0      1.0      0.0                  0.000000
25%        0.0      0.0      0.0      1.0      0.0                  1.000000
50%        0.0      0.0      0.0      1.0      0.0                  2.000000
75%        0.0      0.0      0.0      1.0      0.0                  3.000000
max        0.0      0.0      0.0      1.0      0.0                  4.000000

       White Blood cell count (thousand per microliter)     Symptom 1  \
count                                      19935.000000  19928.000000
mean                                           7.486224      0.592483
std                                            2.653393      0.491385
min                                            3.000000      0.000000
25%                                            5.424703      0.000000
50%                                            7.477132      1.000000
75%                                            9.526152      1.000000
max                                           12.000000      1.000000

          Symptom 2     Symptom 3     Symptom 4     Symptom 5
count  19861.000000  19982.000000  19970.000000  19930.000000
mean       0.551886      0.536233      0.497747      0.461917
std        0.497313      0.498698      0.500007      0.498560
min        0.000000      0.000000      0.000000      0.000000
25%        0.000000      0.000000      0.000000      0.000000
50%        1.000000      1.000000      0.000000      0.000000
75%        1.000000      1.000000      1.000000      1.000000
max        1.000000      1.000000      1.000000      1.000000
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22083 entries, 0 to 22082
Data columns (total 45 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
```

```
 0   Patient Id                                          22083 non-null  object
 1   Patient Age                                         20656 non-null  float64
 2   Genes in mother's side                              22083 non-null  object
 3   Inherited from father                               21777 non-null  object
 4   Maternal gene                                       19273 non-null  object
 5   Paternal gene                                       22083 non-null  object
 6   Blood cell count (mcL)                              22083 non-null  float64
 7   Patient First Name                                  22083 non-null  object
 8   Family Name                                         12392 non-null  object
 9   Father's name                                       22083 non-null  object
 10  Mother's age                                        16047 non-null  float64
 11  Father's age                                        16097 non-null  float64
 12  Institute Name                                      16977 non-null  object
 13  Location of Institute                               22083 non-null  object
 14  Status                                              22083 non-null  object
 15  Respiratory Rate (breaths/min)                      19934 non-null  object
 16  Heart Rate (rates/min                               19970 non-null  object
 17  Test 1                                              19956 non-null  float64
 18  Test 2                                              19931 non-null  float64
 19  Test 3                                              19936 non-null  float64
 20  Test 4                                              19943 non-null  float64
 21  Test 5                                              19913 non-null  float64
 22  Parental consent                                    19958 non-null  object
 23  Follow-up                                           19917 non-null  object
 24  Gender                                              19910 non-null  object
 25  Birth asphyxia                                      19944 non-null  object
 26  Autopsy shows birth defect (if applicable)          17691 non-null  object
 27  Place of birth                                      19959 non-null  object
 28  Folic acid details (peri-conceptional)              19966 non-null  object
 29  H/O serious maternal illness                        19931 non-null  object
 30  H/O radiation exposure (x-ray)                      19930 non-null  object
 31  H/O substance abuse                                 19888 non-null  object
 32  Assisted conception IVF/ART                         19961 non-null  object
 33  History of anomalies in previous pregnancies        19911 non-null  object
 34  No. of previous abortion                            19921 non-null  float64
 35  Birth defects                                       19929 non-null  object
 36  White Blood cell count (thousand per microliter)    19935 non-null  float64
 37  Blood test result                                   19938 non-null  object
 38  Symptom 1                                           19928 non-null  float64
 39  Symptom 2                                           19861 non-null  float64
 40  Symptom 3                                           19982 non-null  float64
 41  Symptom 4                                           19970 non-null  float64
 42  Symptom 5                                           19930 non-null  float64
 43  Genetic Disorder                                    19937 non-null  object
 44  Disorder Subclass                                   19915 non-null  object
dtypes: float64(16), object(29)
memory usage: 7.6+ MB
```

[6]:
```
df = df.drop(columns=["Patient Id", "Patient First Name", "Family Name",␣
↪"Father's name"],  axis=1)
```

[7]:
```
df.columns
```

[7]:
```
Index(['Patient Age', 'Genes in mother's side', 'Inherited from father',
       'Maternal gene', 'Paternal gene', 'Blood cell count (mcL)',
       'Mother's age', 'Father's age', 'Institute Name',
       'Location of Institute', 'Status', 'Respiratory Rate (breaths/min)',
       'Heart Rate (rates/min', 'Test 1', 'Test 2', 'Test 3', 'Test 4',
       'Test 5', 'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

[8]:
```
df["Genes in mother's side"].head()
```

[8]:
```
0    Yes
1    Yes
2    Yes
3    Yes
4    Yes
Name: Genes in mother's side, dtype: object
```

[9]:
```
num = df.select_dtypes(include=np.number)
cat = df.select_dtypes(exclude=np.number)
```

[10]:
```
df.isnull().sum()
```

[10]:
```
Patient Age                                1427
Genes in mother's side                        0
Inherited from father                       306
Maternal gene                              2810
Paternal gene                                 0
Blood cell count (mcL)                        0
Mother's age                               6036
Father's age                               5986
Institute Name                             5106
Location of Institute                         0
Status                                        0
```

```
Respiratory Rate (breaths/min)                    2149
Heart Rate (rates/min                             2113
Test 1                                            2127
Test 2                                            2152
Test 3                                            2147
Test 4                                            2140
Test 5                                            2170
Parental consent                                  2125
Follow-up                                         2166
Gender                                            2173
Birth asphyxia                                    2139
Autopsy shows birth defect (if applicable)        4392
Place of birth                                    2124
Folic acid details (peri-conceptional)            2117
H/O serious maternal illness                      2152
H/O radiation exposure (x-ray)                    2153
H/O substance abuse                               2195
Assisted conception IVF/ART                       2122
History of anomalies in previous pregnancies      2172
No. of previous abortion                          2162
Birth defects                                     2154
White Blood cell count (thousand per microliter)  2148
Blood test result                                 2145
Symptom 1                                         2155
Symptom 2                                         2222
Symptom 3                                         2101
Symptom 4                                         2113
Symptom 5                                         2153
Genetic Disorder                                  2146
Disorder Subclass                                 2168
dtype: int64
```

[11]:
```
num.columns
```

[11]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)', 'Symptom 1',
       'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5'],
      dtype='object')
```

[12]:
```
num.isnull().sum()
```

[12]:
```
Patient Age                                       1427
Blood cell count (mcL)                               0
Mother's age                                      6036
Father's age                                      5986
```

```
Test 1                                                2127
Test 2                                                2152
Test 3                                                2147
Test 4                                                2140
Test 5                                                2170
No. of previous abortion                              2162
White Blood cell count (thousand per microliter)      2148
Symptom 1                                             2155
Symptom 2                                             2222
Symptom 3                                             2101
Symptom 4                                             2113
Symptom 5                                             2153
dtype: int64
```

[13]:
```
num.fillna(num.mean(), inplace=True)
num.isnull().sum()
```

[13]:
```
Patient Age                                           0
Blood cell count (mcL)                                0
Mother's age                                          0
Father's age                                          0
Test 1                                                0
Test 2                                                0
Test 3                                                0
Test 4                                                0
Test 5                                                0
No. of previous abortion                              0
White Blood cell count (thousand per microliter)      0
Symptom 1                                             0
Symptom 2                                             0
Symptom 3                                             0
Symptom 4                                             0
Symptom 5                                             0
dtype: int64
```

[14]:
```
cat.columns
```

[14]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
```

```
      dtype='object')
```

[15]:
```
cat.fillna(cat.mode(), inplace=True)
cat.isnull().sum()
```

[15]:
```
Genes in mother's side                            0
Inherited from father                           306
Maternal gene                                  2810
Paternal gene                                     0
Institute Name                                 5106
Location of Institute                             0
Status                                            0
Respiratory Rate (breaths/min)                 2149
Heart Rate (rates/min                          2113
Parental consent                               2125
Follow-up                                      2166
Gender                                         2172
Birth asphyxia                                 2138
Autopsy shows birth defect (if applicable)     4392
Place of birth                                 2124
Folic acid details (peri-conceptional)         2117
H/O serious maternal illness                   2151
H/O radiation exposure (x-ray)                 2153
H/O substance abuse                            2195
Assisted conception IVF/ART                    2122
History of anomalies in previous pregnancies   2172
Birth defects                                  2153
Blood test result                              2144
Genetic Disorder                               2146
Disorder Subclass                              2168
dtype: int64
```

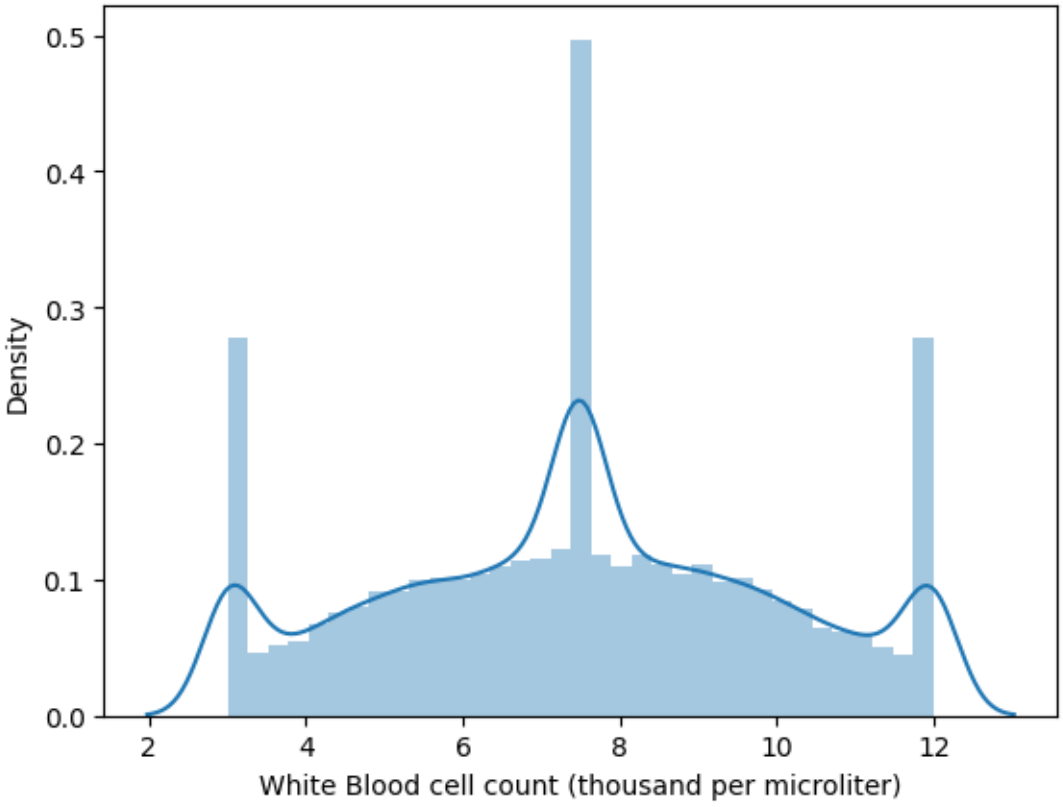
[16]:
```
for column in cat.columns:
 cat[column].fillna(cat[column].mode()[0], inplace=True)
cat.isnull().sum()
```

[16]:
```
Genes in mother's side                    0
Inherited from father                     0
Maternal gene                             0
Paternal gene                             0
Institute Name                            0
Location of Institute                     0
Status                                    0
Respiratory Rate (breaths/min)            0
Heart Rate (rates/min                     0
Parental consent                          0
Follow-up                                 0
```

```
Gender                                          0
Birth asphyxia                                  0
Autopsy shows birth defect (if applicable)      0
Place of birth                                  0
Folic acid details (peri-conceptional)          0
H/O serious maternal illness                    0
H/O radiation exposure (x-ray)                  0
H/O substance abuse                             0
Assisted conception IVF/ART                     0
History of anomalies in previous pregnancies    0
Birth defects                                   0
Blood test result                               0
Genetic Disorder                                0
Disorder Subclass                               0
dtype: int64
```

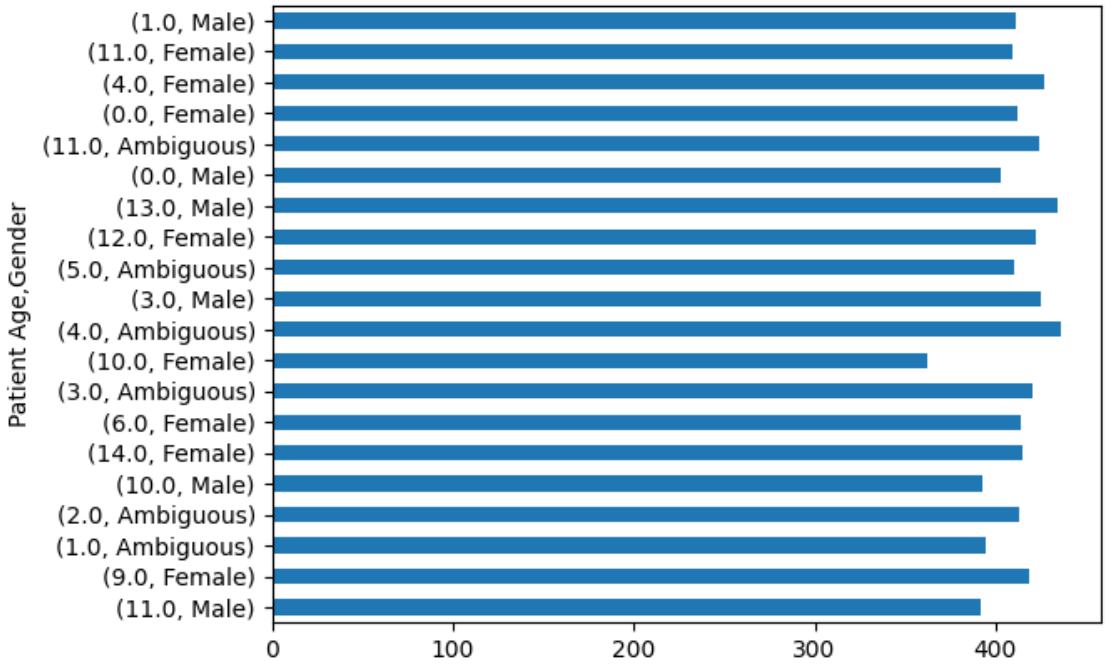
[17]:
```
sns.distplot(num['White Blood cell count (thousand per microliter)'])
```

[17]:
```
<Axes: xlabel='White Blood cell count (thousand per microliter)',
ylabel='Density'>
```



[19]:
```
#age and gender
ag = df.groupby('Patient Age')['Gender'].value_counts()
ag

ag2 = ag.sample(20)

ag2.plot(kind='barh',)
```

[19]:
```
<Axes: ylabel='Patient Age,Gender'>
```

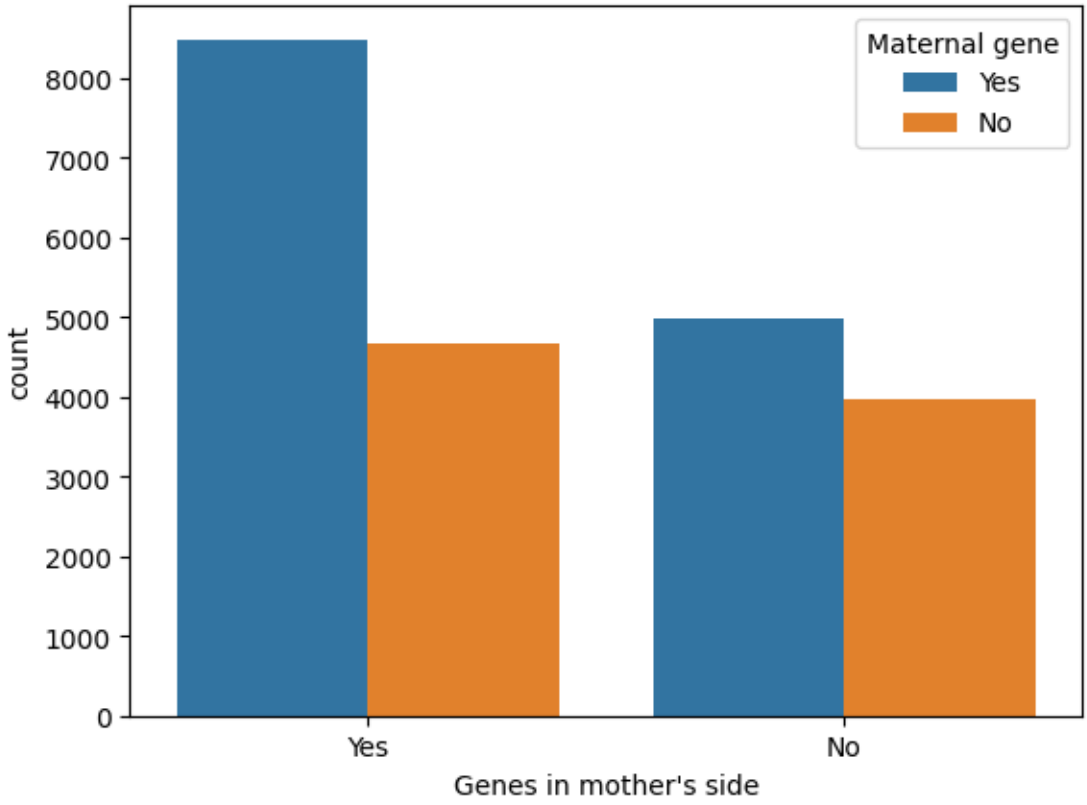


[20]:
```
cat.columns
```

[20]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

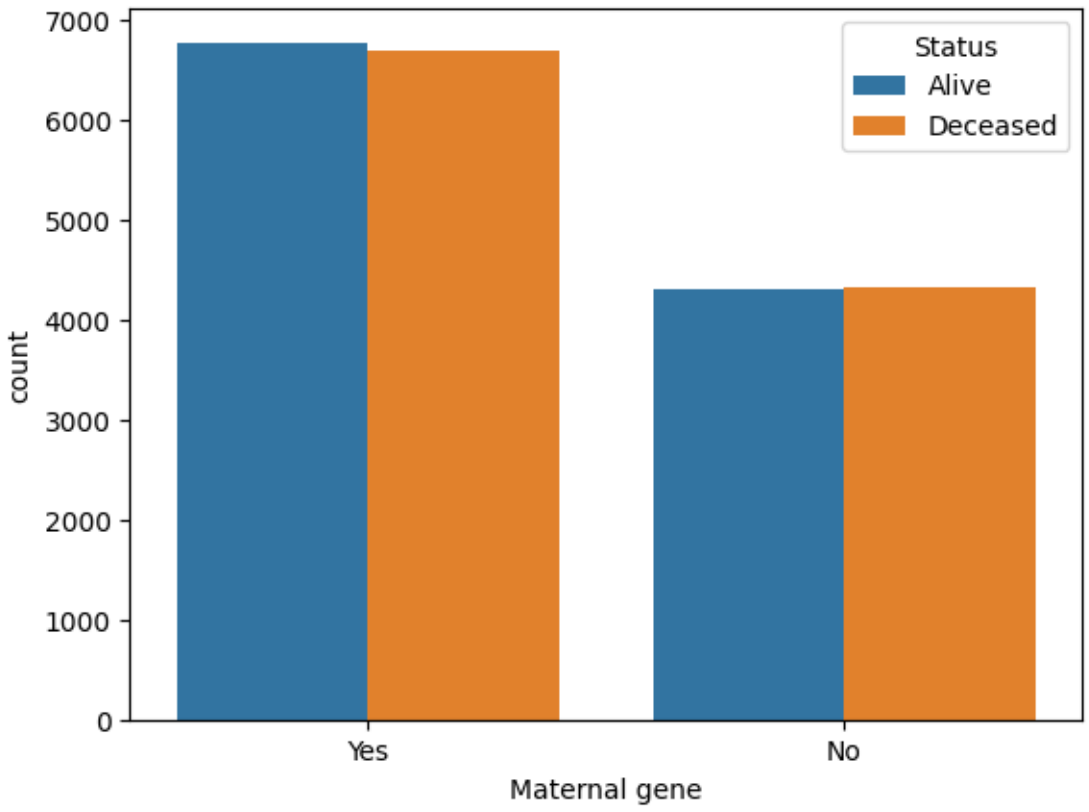
[21]:
```
sns.countplot(x="Genes in mother's side", data=cat, hue="Maternal gene")
```

[21]: <Axes: xlabel="Genes in mother's side", ylabel='count'>



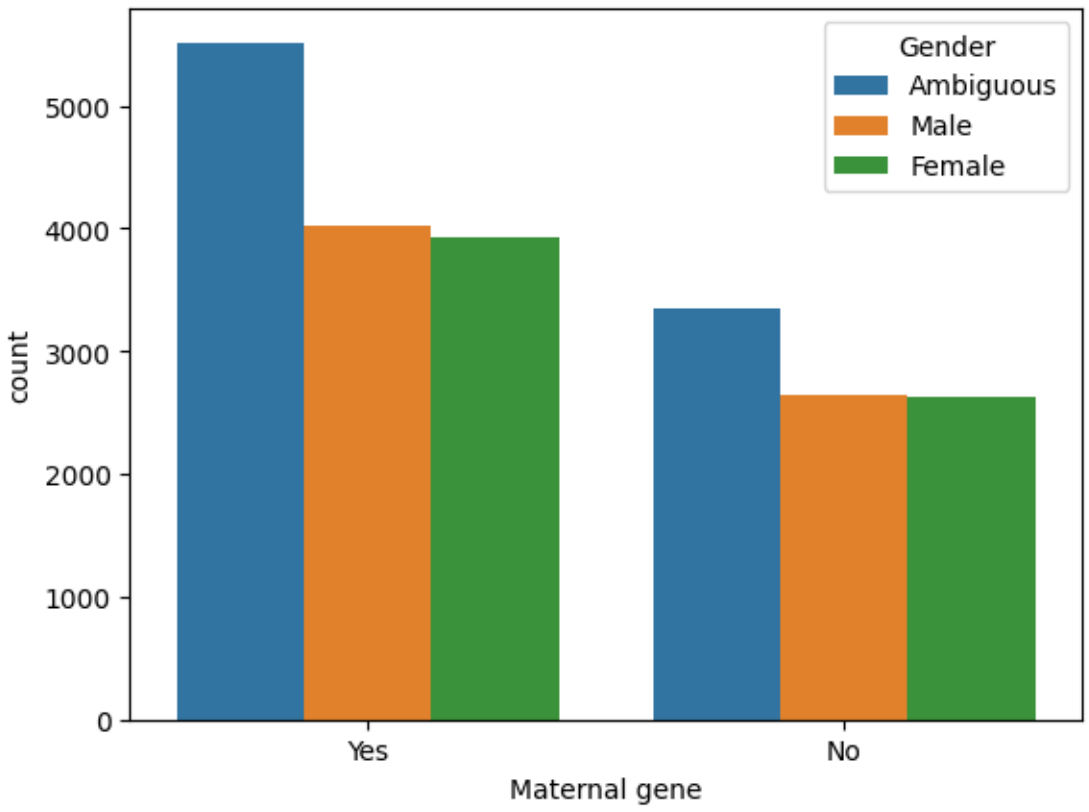
[22]:
```
sns.countplot(x="Maternal gene", data=cat, hue="Status")
```

[22]: <Axes: xlabel='Maternal gene', ylabel='count'>



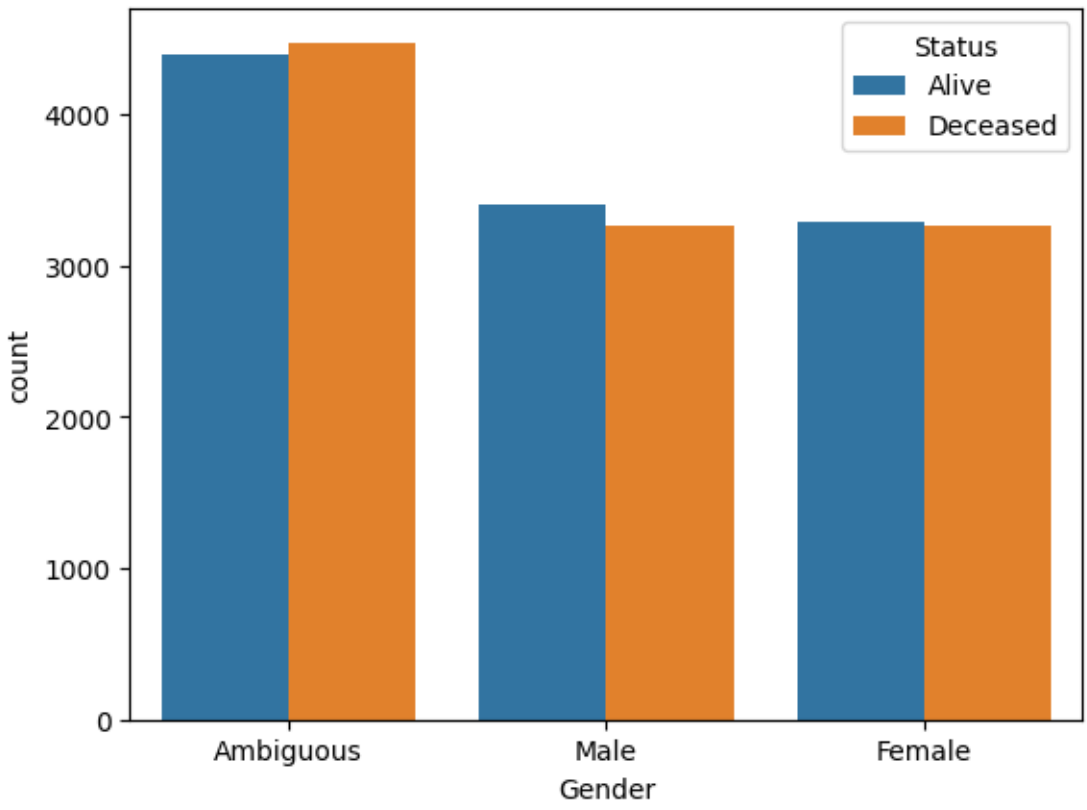
[23]:
```
sns.countplot(x="Maternal gene", data=cat, hue="Gender")
```

[23]:
```
<Axes: xlabel='Maternal gene', ylabel='count'>
```



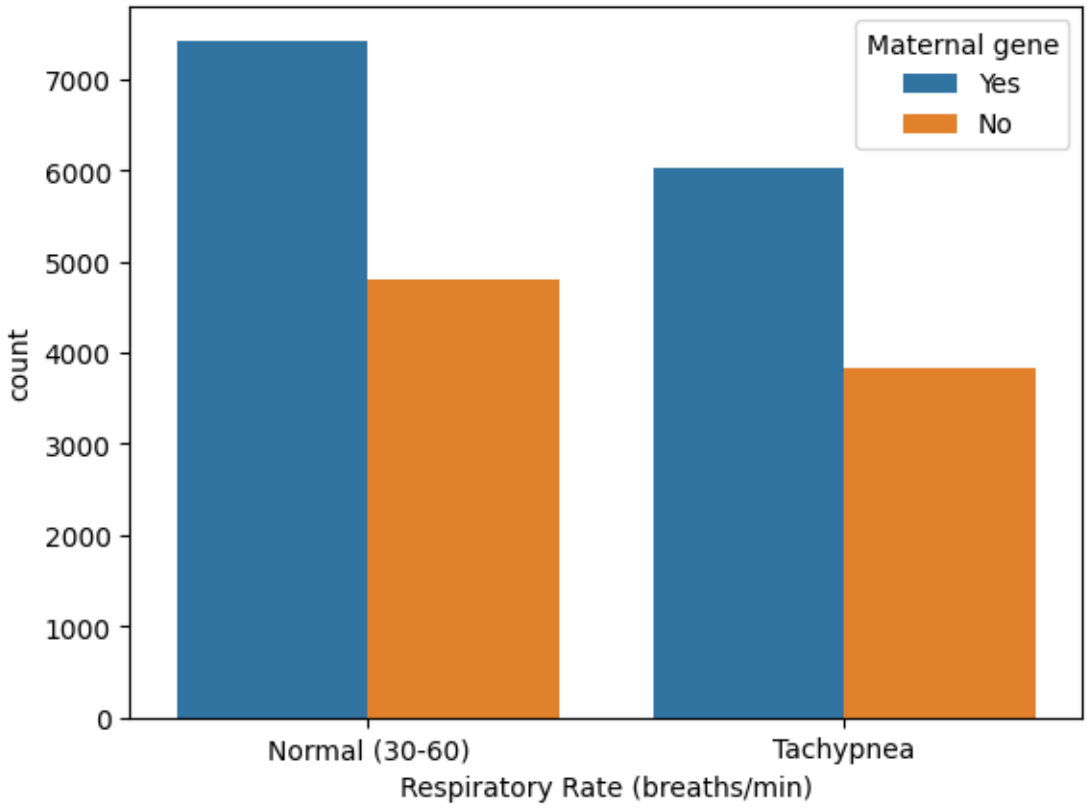
[24]:
```
sns.countplot(x="Gender", data=cat, hue="Status")
```

[24]:
```
<Axes: xlabel='Gender', ylabel='count'>
```



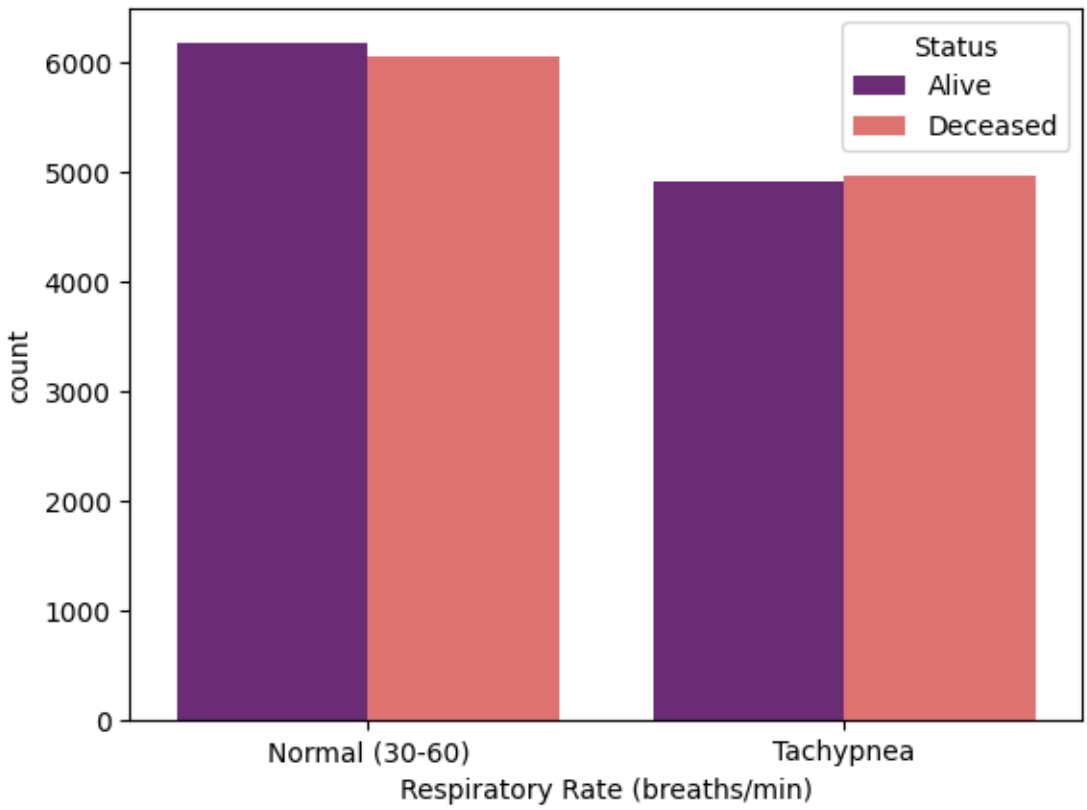
[25]:
```
sns.countplot(x="Respiratory Rate (breaths/min)", data=cat, hue="Maternal gene")
```

[25]:
```
<Axes: xlabel='Respiratory Rate (breaths/min)', ylabel='count'>
```



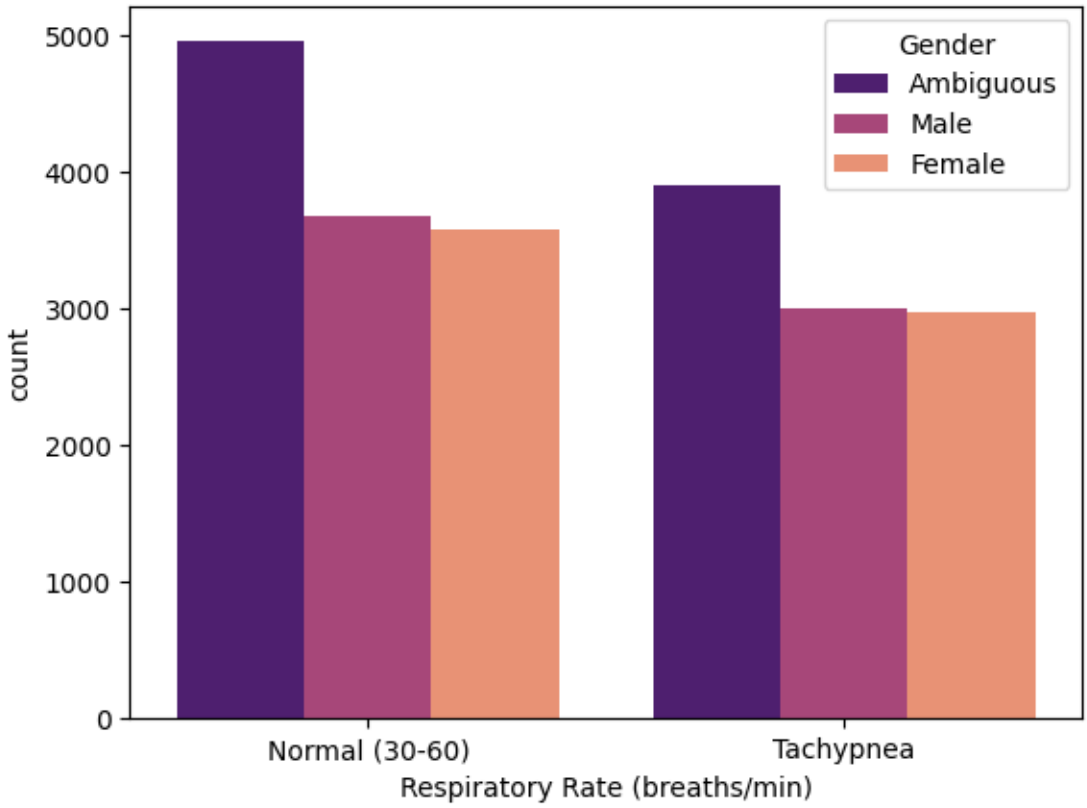
[26]:
```
sns.countplot(x="Respiratory Rate (breaths/min)", data=cat, hue="Status",␣
↪palette='magma')
```

[26]:
```
<Axes: xlabel='Respiratory Rate (breaths/min)', ylabel='count'>
```



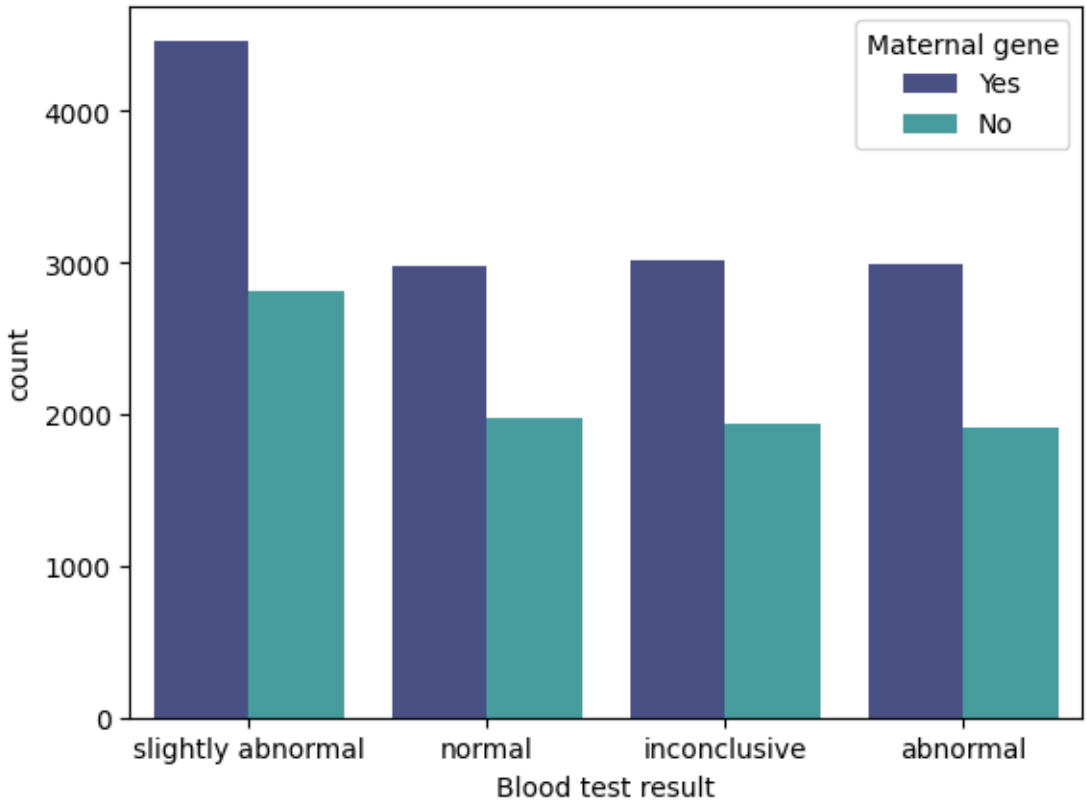
[27]:
```
sns.countplot(x="Respiratory Rate (breaths/min)", data=cat,␣
↪hue="Gender",palette='magma')
```

[27]:
```
<Axes: xlabel='Respiratory Rate (breaths/min)', ylabel='count'>
```



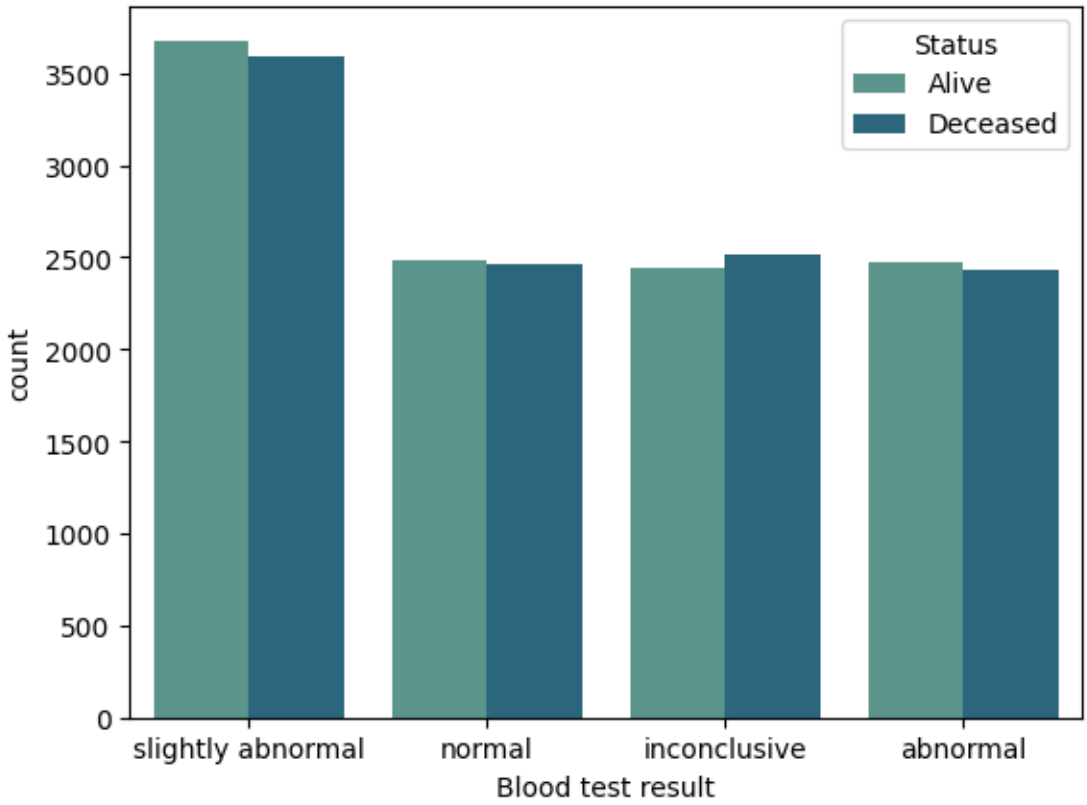
[28]:
```
sns.countplot(x="Blood test result", data=cat, hue="Maternal gene",␣
↪palette='mako')
```

[28]:
```
<Axes: xlabel='Blood test result', ylabel='count'>
```



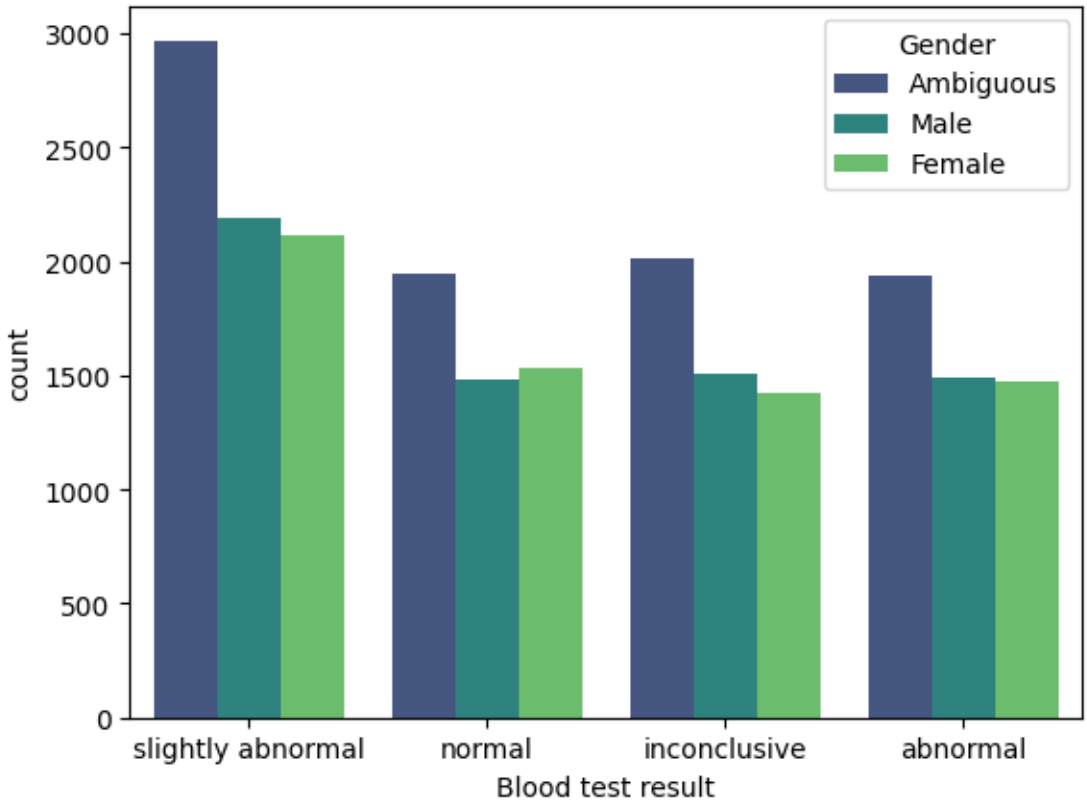
[29]:
```
sns.countplot(x="Blood test result", data=cat, hue="Status", palette='crest')
```

[29]:
```
<Axes: xlabel='Blood test result', ylabel='count'>
```



[30]:
```
sns.countplot(x="Blood test result", data=cat, hue="Gender", palette =␣
↪'viridis')
```

[30]:
```
<Axes: xlabel='Blood test result', ylabel='count'>
```

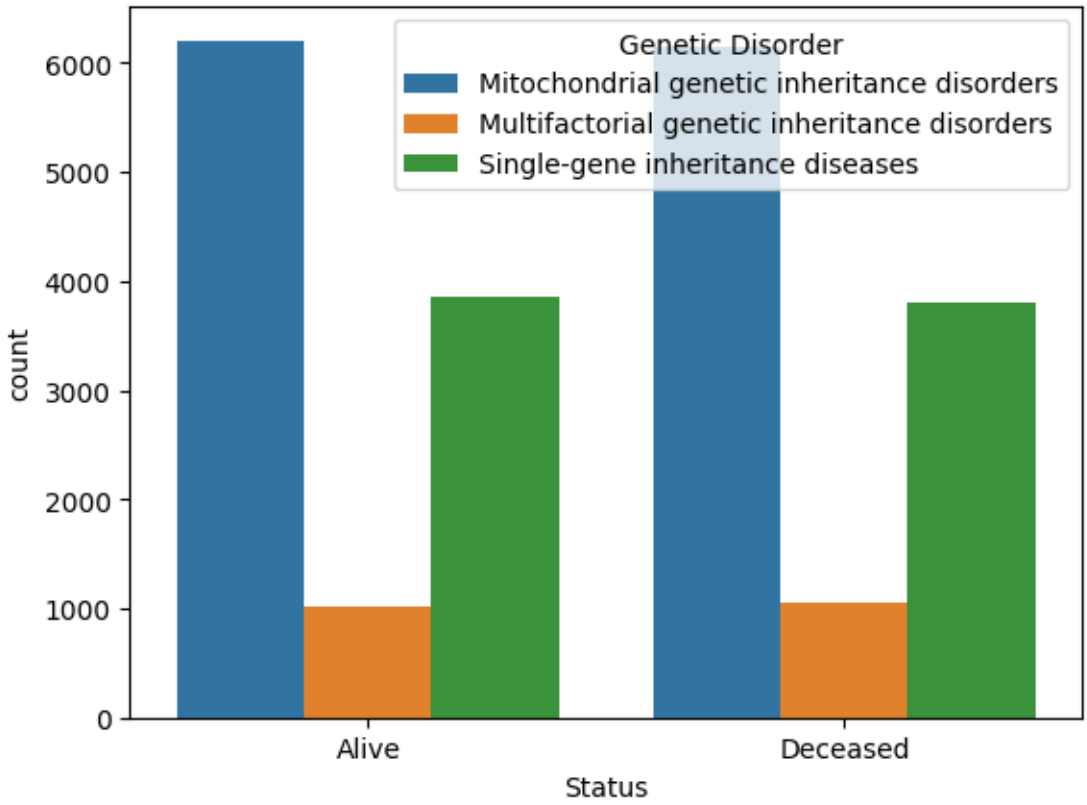


[31]:
```
cat.columns
```

[31]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

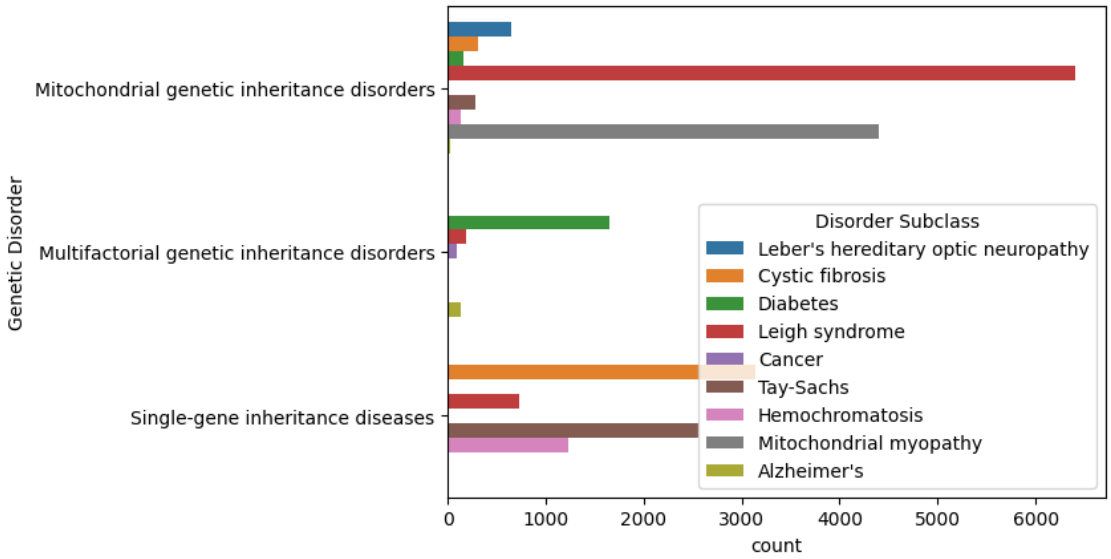
[32]:
```
sns.countplot(x="Status", data=cat, hue="Genetic Disorder")
```

[32]:
```
<Axes: xlabel='Status', ylabel='count'>
```



[33]:
```
sns.countplot(y="Genetic Disorder", data=cat, hue="Disorder Subclass")
```

[33]: <Axes: xlabel='count', ylabel='Genetic Disorder'>

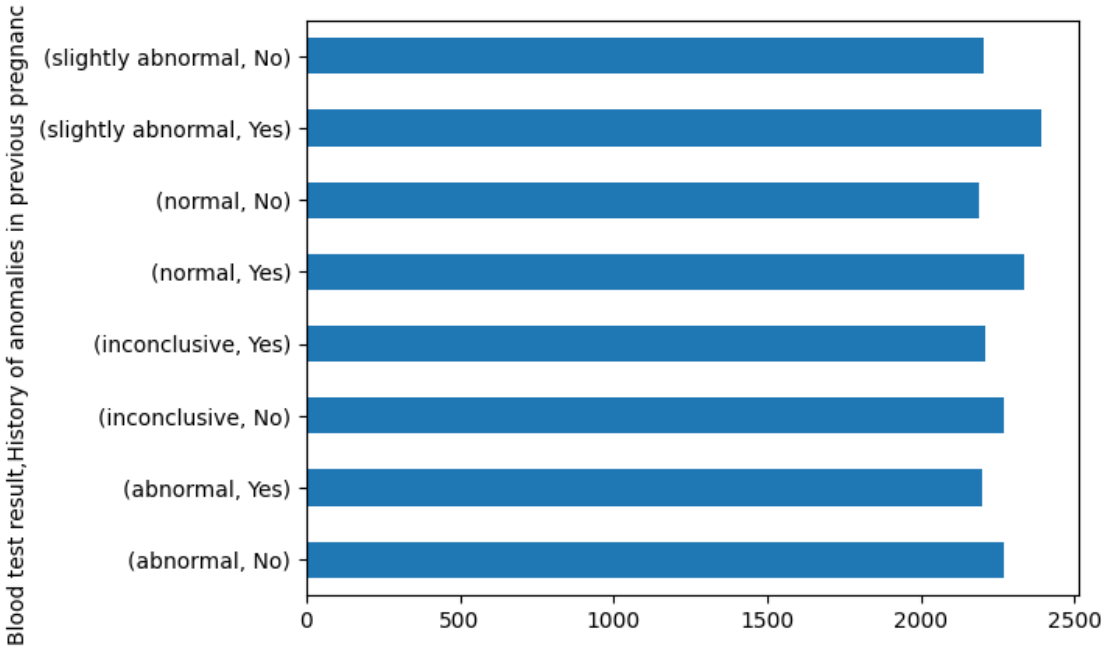


[34]:
```
mental = df['H/O serious maternal illness']
mental.head()
mental.dropna(inplace=True)
mental.head()
```

[34]:
```
1    Yes
2     No
3    Yes
4    Yes
5     No
Name: H/O serious maternal illness, dtype: object
```

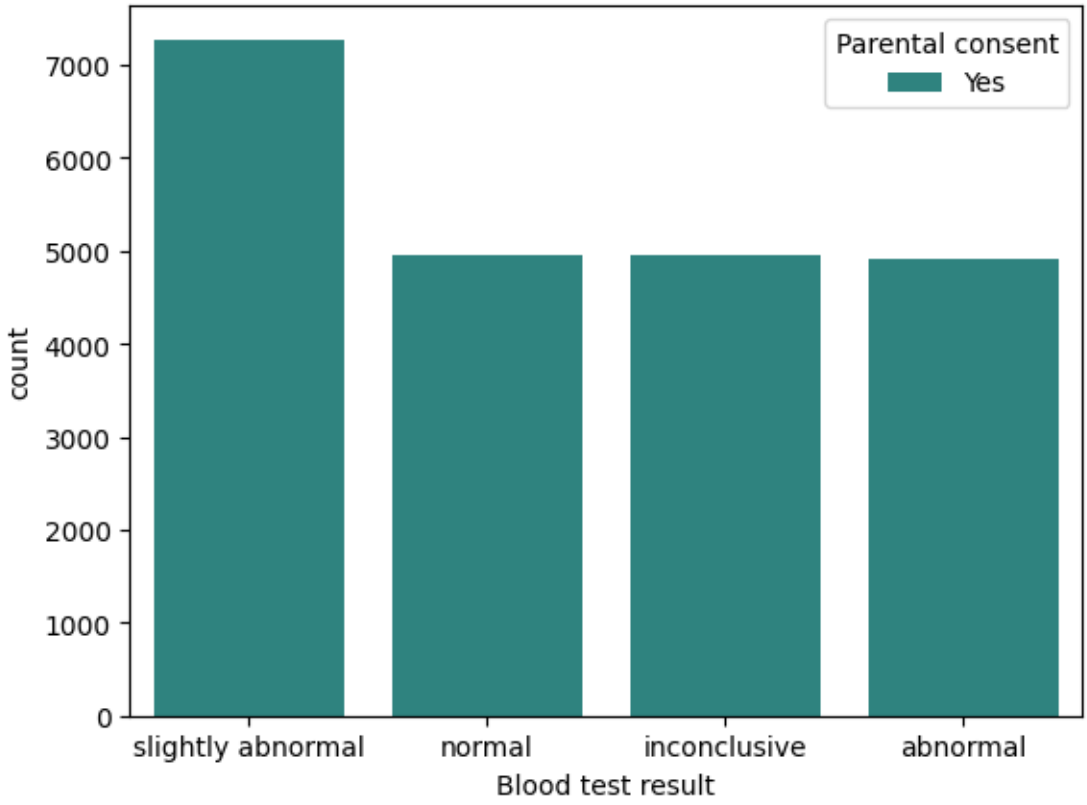
[35]:
```
# blood test results and pregnancies

bp = df.groupby('Blood test result')['History of anomalies in previous␣
 ↪pregnancies'].value_counts()
bp

bp.plot(kind='barh')
```

[35]:
```
<Axes: ylabel='Blood test result,History of anomalies in previous pregnancies'>
```



[36]:
```
sns.countplot(x="Blood test result", data=cat, hue="Parental consent",␣
 ↪palette='viridis')
```

[36]: <Axes: xlabel='Blood test result', ylabel='count'>

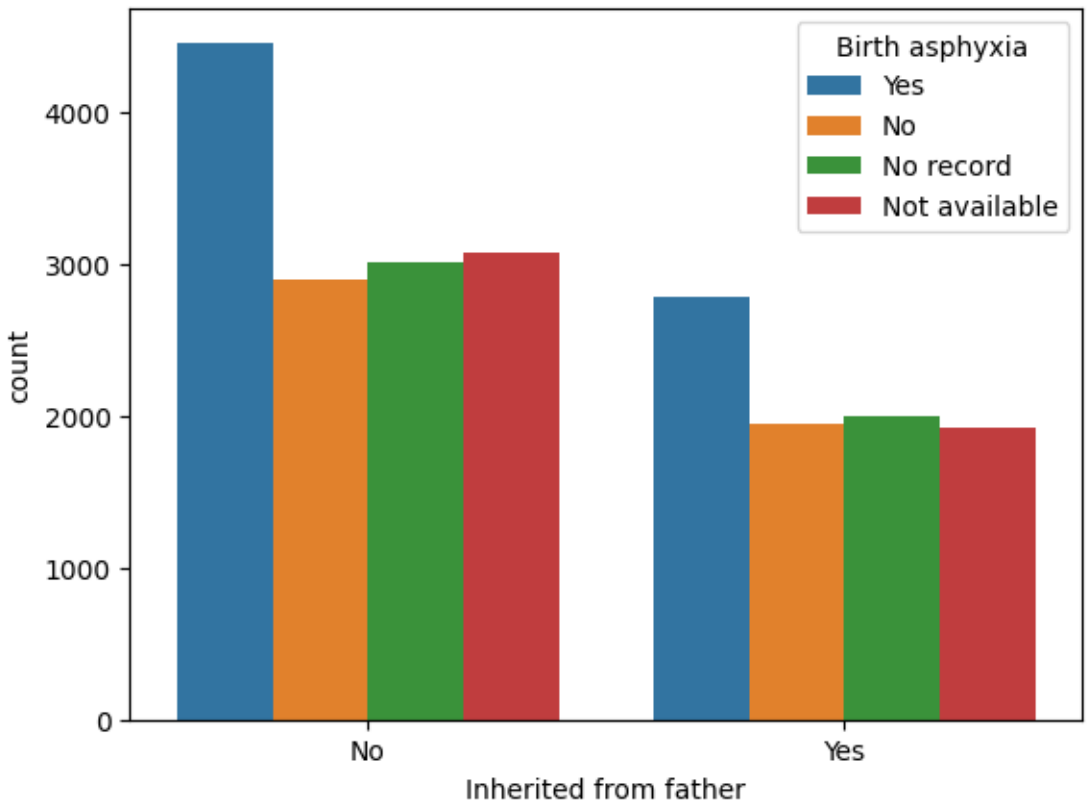


[37]:
```
cat.columns
```

[37]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

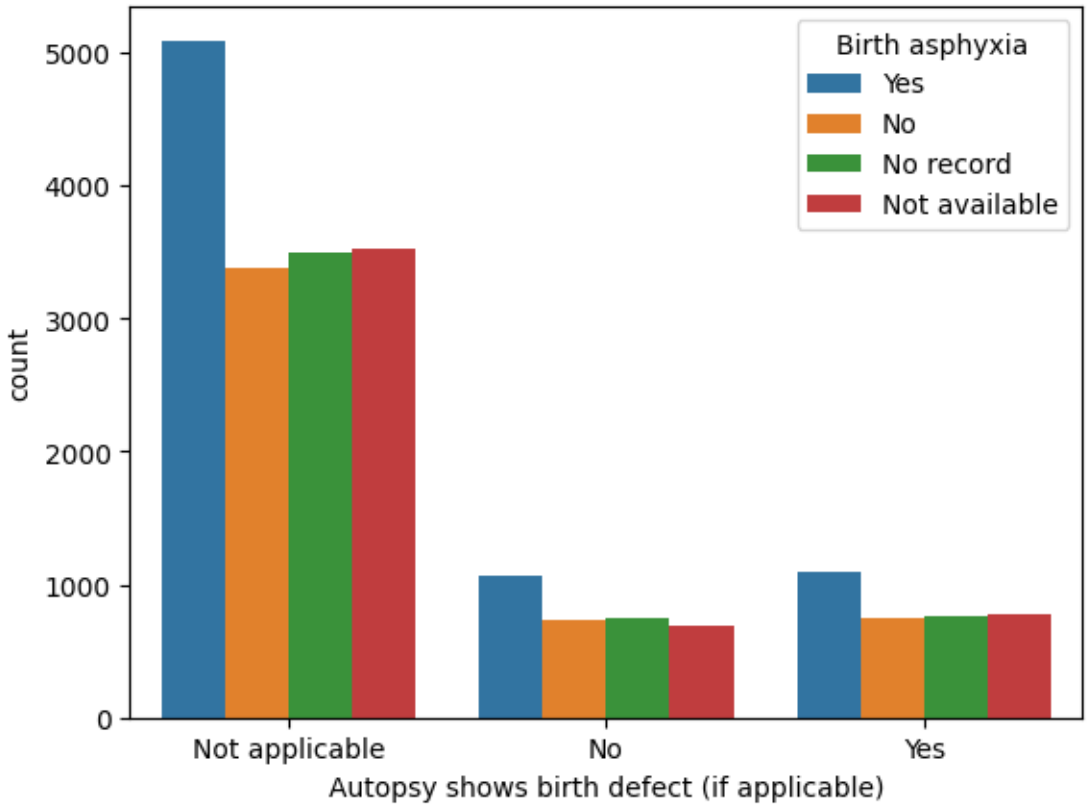
[38]:
```
sns.countplot(x="Inherited from father", hue="Birth asphyxia", data=cat)
```

[38]: <Axes: xlabel='Inherited from father', ylabel='count'>



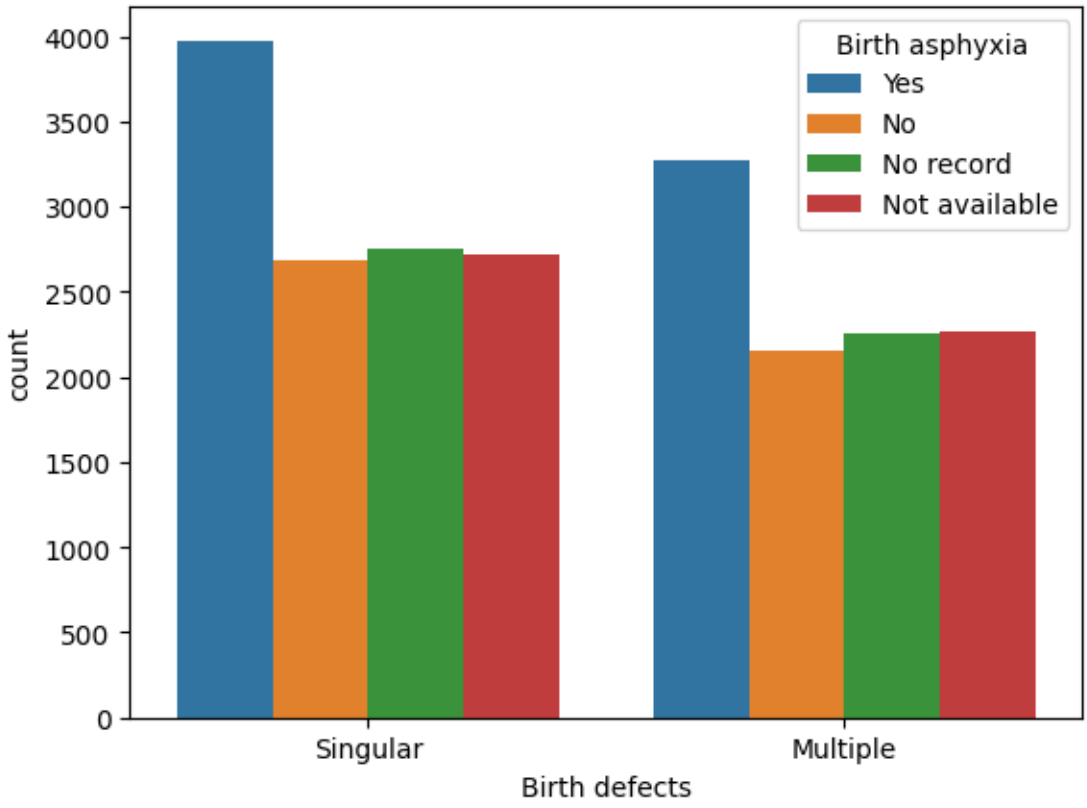
[39]:
```
sns.countplot(x="Autopsy shows birth defect (if applicable)", hue="Birth␣
↪asphyxia", data=cat)
```

[39]: <Axes: xlabel='Autopsy shows birth defect (if applicable)', ylabel='count'>



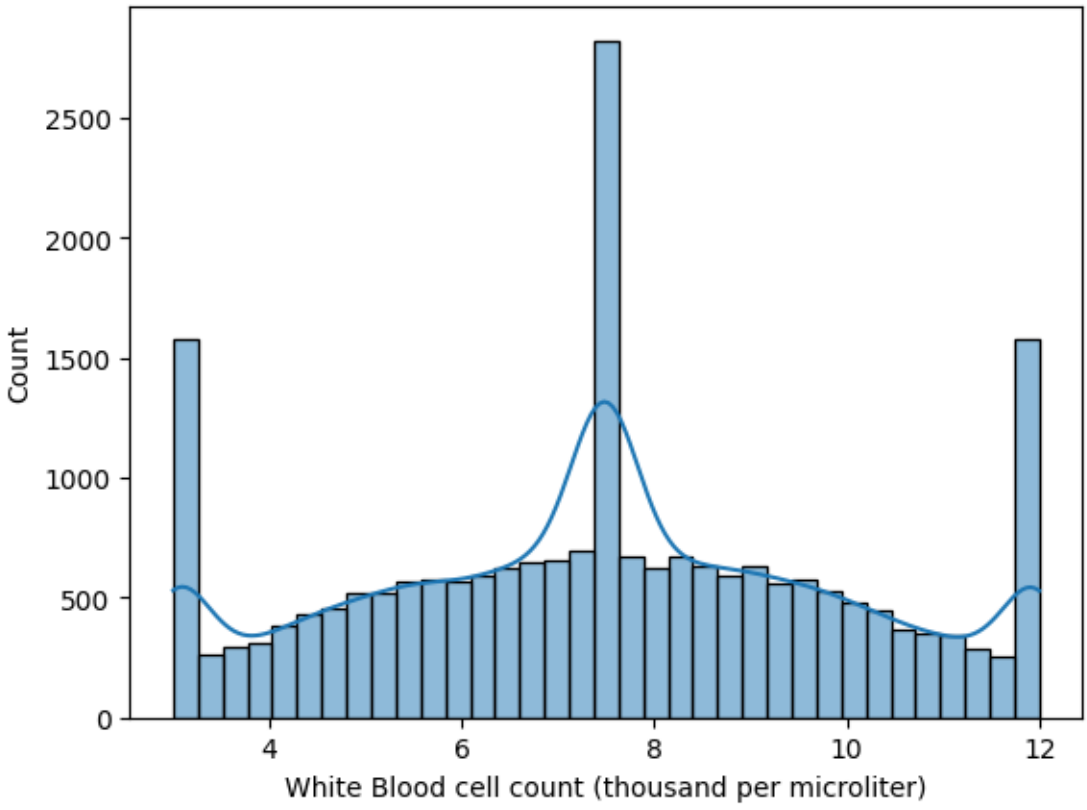
[40]:
```
sns.countplot(x="Birth defects", hue="Birth asphyxia", data=cat)
```

[40]:
```
<Axes: xlabel='Birth defects', ylabel='count'>
```



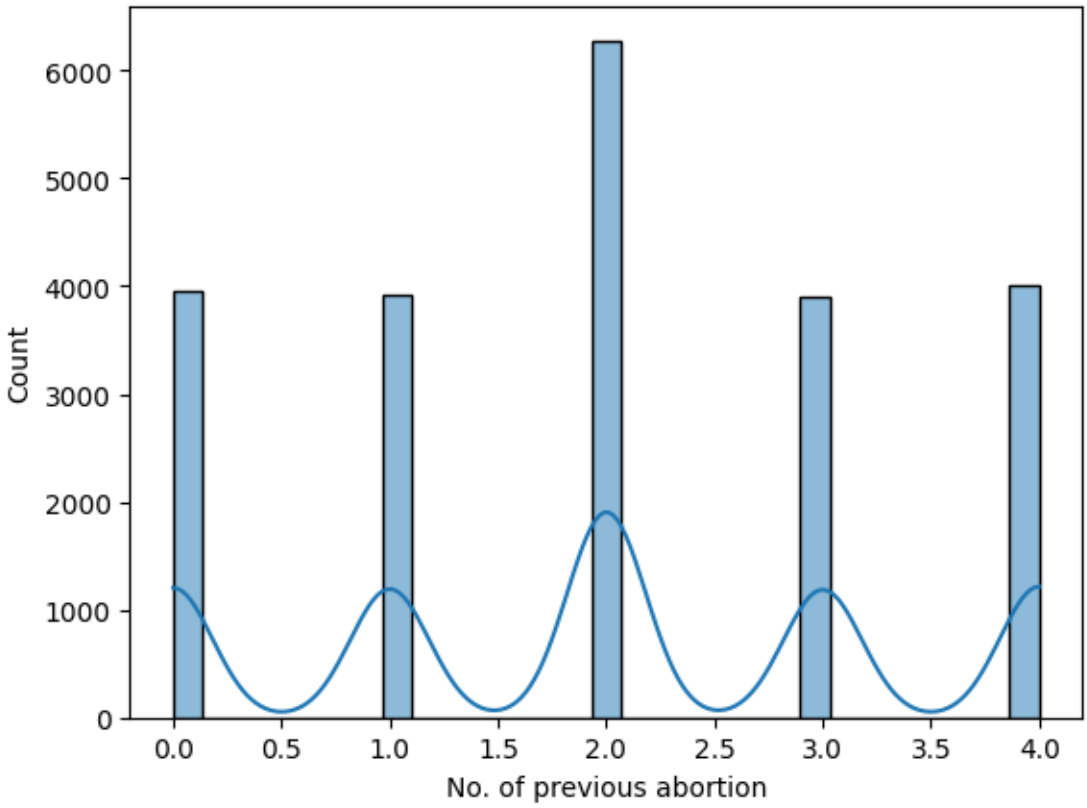
[41]:
```
sns.histplot(x='White Blood cell count (thousand per microliter)', kde=True,␣
↪data=num)
```

[41]:
```
<Axes: xlabel='White Blood cell count (thousand per microliter)',
ylabel='Count'>
```



[42]:
```
sns.histplot(x='No. of previous abortion', kde=True, data=num)
```

[42]:
```
<Axes: xlabel='No. of previous abortion', ylabel='Count'>
```



[43]:
```
cat.columns
```

[43]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

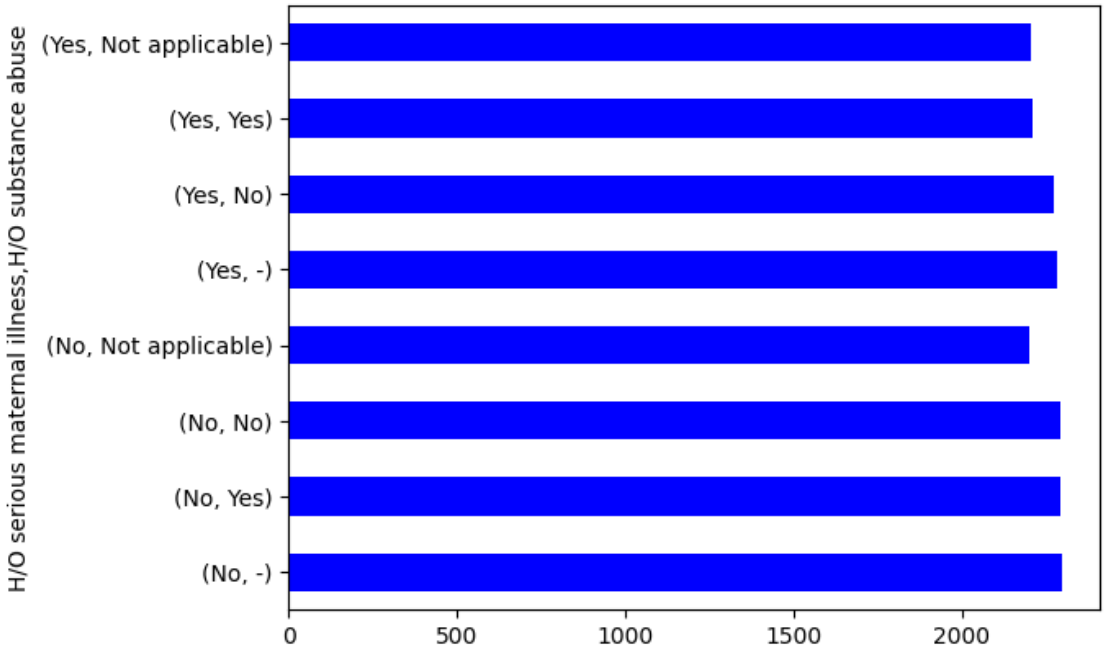
[44]:
```
num.columns
```

[44]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)', 'Symptom 1',
       'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5'],
```

```
        dtype='object')
```

[45]:
```
# serious mental illness vs substance use

ss = df.groupby('H/O serious maternal illness')['H/O substance abuse'].
 ↪value_counts()
ss

ss.plot(kind='barh', color='blue')
```

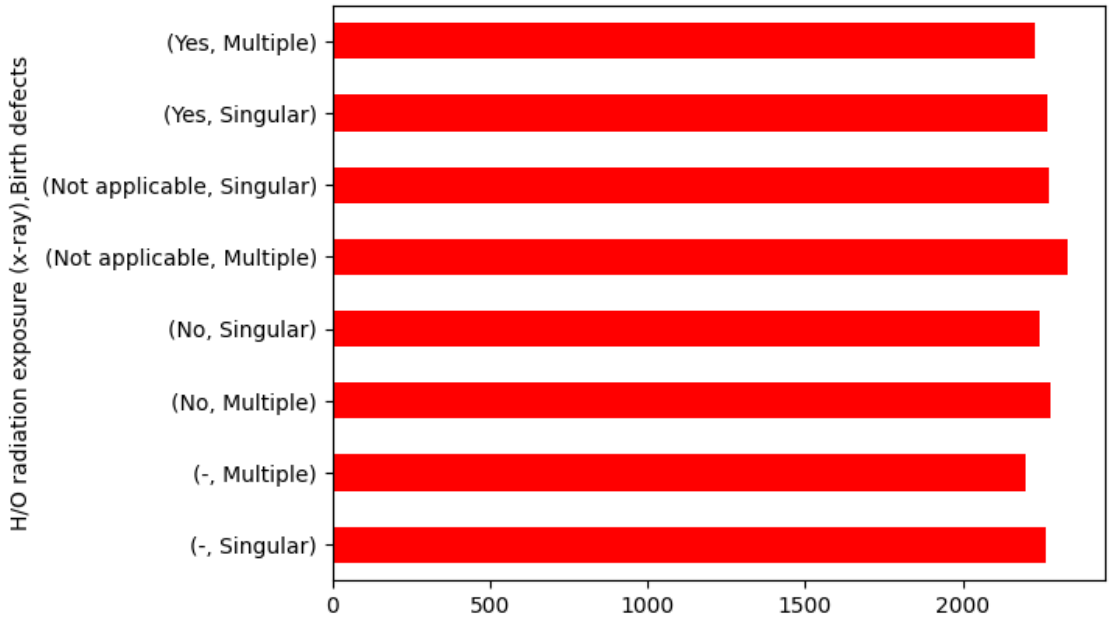
[45]:
```
<Axes: ylabel='H/O serious maternal illness,H/O substance abuse'>
```



[46]:
```
columns = df.columns
for col in columns:
  if "Not applicable" in col:
    df.drop(columns=col, inplace=True)
```

[47]:
```
# radiation exposure and birth defects

rb = df.groupby('H/O radiation exposure (x-ray)')['Birth defects'].
 ↪value_counts()
rb

rb.plot(kind='barh', color='red')
```

[47]: <Axes: ylabel='H/O radiation exposure (x-ray),Birth defects'>



[48]:
```
rad = df['H/O radiation exposure (x-ray)']
rad.isna().sum()
```

[48]: 2153

[49]:
```
df['Birth defects'].value_counts()
```

[49]:
```
Birth defects
Singular    9977
Multiple    9952
Name: count, dtype: int64
```

[50]:
```
df.columns
```

[50]:
```
Index(['Patient Age', 'Genes in mother's side', 'Inherited from father',
       'Maternal gene', 'Paternal gene', 'Blood cell count (mcL)',
       'Mother's age', 'Father's age', 'Institute Name',
       'Location of Institute', 'Status', 'Respiratory Rate (breaths/min)',
       'Heart Rate (rates/min', 'Test 1', 'Test 2', 'Test 3', 'Test 4',
       'Test 5', 'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
```

```
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

[51]:
```
num.columns
```

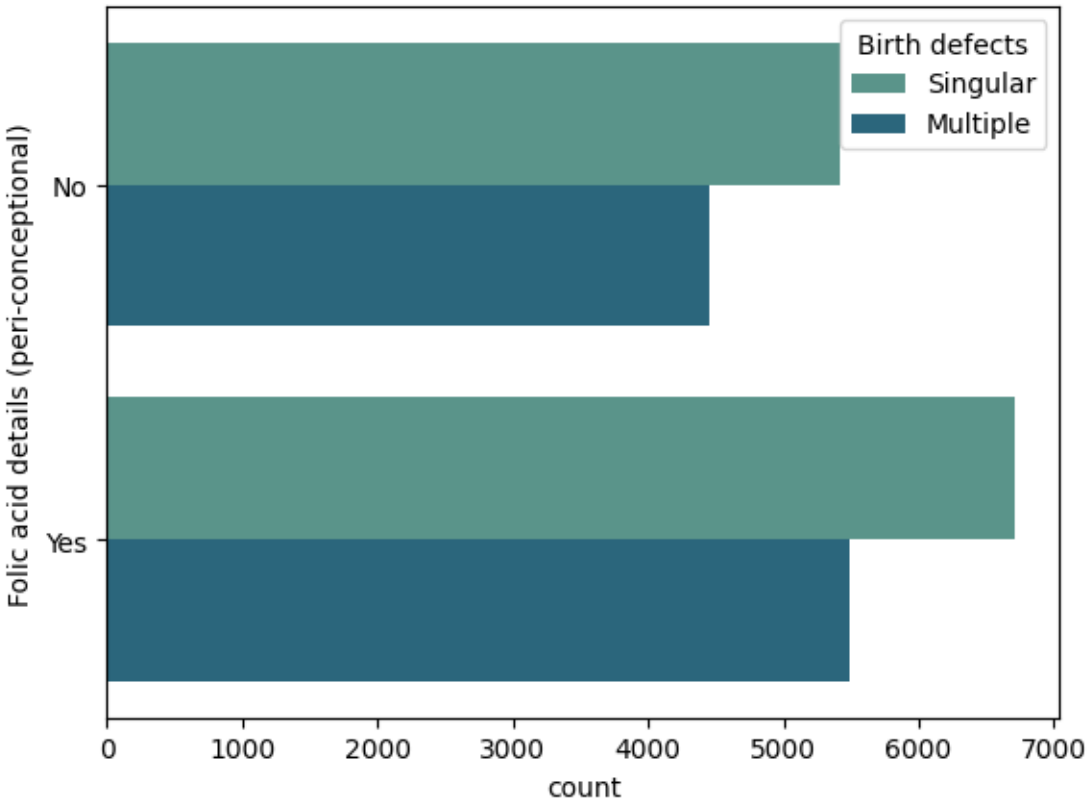
[51]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)', 'Symptom 1',
       'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5'],
      dtype='object')
```

[52]:
```
cat.columns
```

[52]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

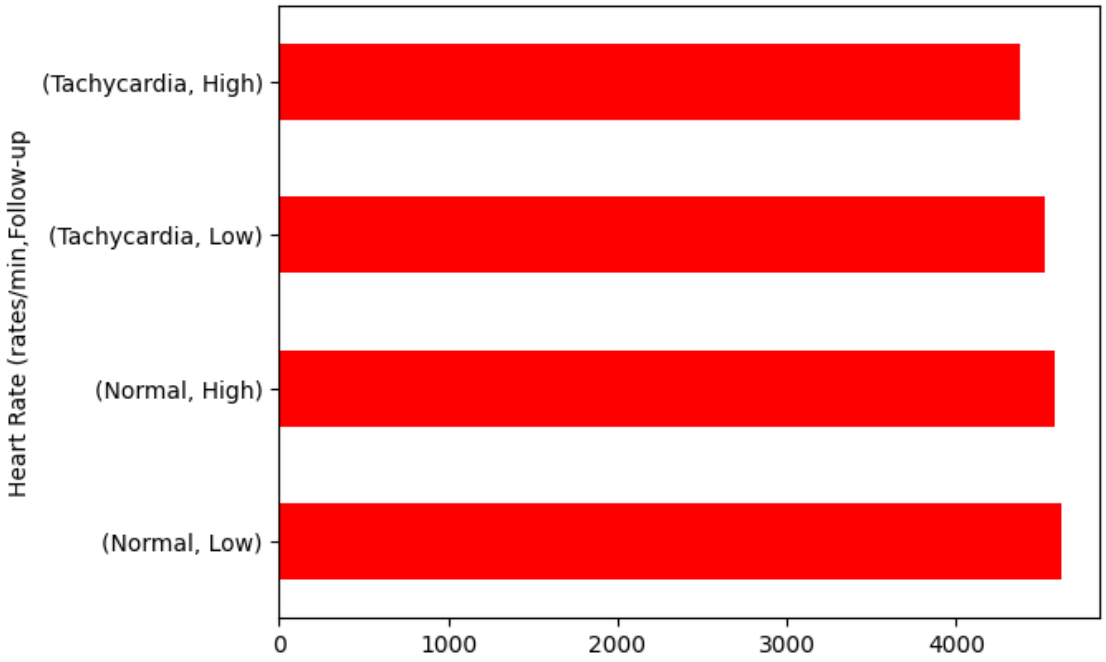
[53]:
```
sns.countplot(y="Folic acid details (peri-conceptional)", hue="Birth defects",␣
 ↪data=cat, palette='crest')
```

[53]:
```
<Axes: xlabel='count', ylabel='Folic acid details (peri-conceptional)'>
```



[54]:
```
# heart rate, follow up

hf = df.groupby('Heart Rate (rates/min')['Follow-up'].value_counts()
hf

hf.plot(kind='barh', color='red')
```

[54]: <Axes: ylabel='Heart Rate (rates/min,Follow-up'>



[55]:
```
tests = df[['Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5']]
tests.head()
```

[55]:

	Test 1	Test 2	Test 3	Test 4	Test 5
0	0.0	NaN	NaN	1.0	0.0
1	NaN	0.0	0.0	1.0	0.0
2	0.0	0.0	0.0	1.0	0.0
3	0.0	0.0	0.0	1.0	0.0
4	0.0	0.0	0.0	1.0	0.0

[56]:
```
tests1 = tests.dropna()
```

[57]:
```
tests2 = tests.sample(50)
tests2.head()
```

[57]:

	Test 1	Test 2	Test 3	Test 4	Test 5
19881	0.0	0.0	0.0	1.0	0.0
3204	0.0	0.0	0.0	1.0	0.0
16775	0.0	0.0	NaN	1.0	0.0
10537	0.0	0.0	0.0	1.0	0.0
14401	0.0	0.0	0.0	1.0	0.0

[58]:
```
genes = df[["Genes in mother's side", "Inherited from father", "Maternal gene",
       "Paternal gene"]]
genes.head()
```

[58]:

	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene
0	Yes	No	Yes	No
1	Yes	Yes	No	No
2	Yes	No	No	No
3	Yes	No	Yes	No
4	Yes	No	NaN	Yes

[59]:
```
symptoms = df[['Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4',
       'Symptom 5']]
symptoms.head()
```

[59]:

	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5
0	1.0	1.0	1.0	1.0	1.0
1	1.0	NaN	1.0	1.0	0.0
2	0.0	1.0	1.0	1.0	1.0
3	0.0	0.0	1.0	0.0	0.0
4	0.0	0.0	0.0	0.0	NaN

[60]:
```
symptoms2 = symptoms.dropna()
symptoms2.head()
```

[60]:

	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5
0	1.0	1.0	1.0	1.0	1.0
2	0.0	1.0	1.0	1.0	1.0
3	0.0	0.0	1.0	0.0	0.0
5	1.0	0.0	0.0	1.0	0.0
6	0.0	0.0	0.0	0.0	0.0

[61]:
```
symptoms2 = symptoms2.sample(50)
```

[62]:
```
num2 = num.sample(100)
num2 = num2.drop(columns=['Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4',␣
 ↪'Symptom 5','Test 1', 'Test 2', 'Test 3', 'Test 4',
       'Test 5'], errors='ignore')
num2.head()
```

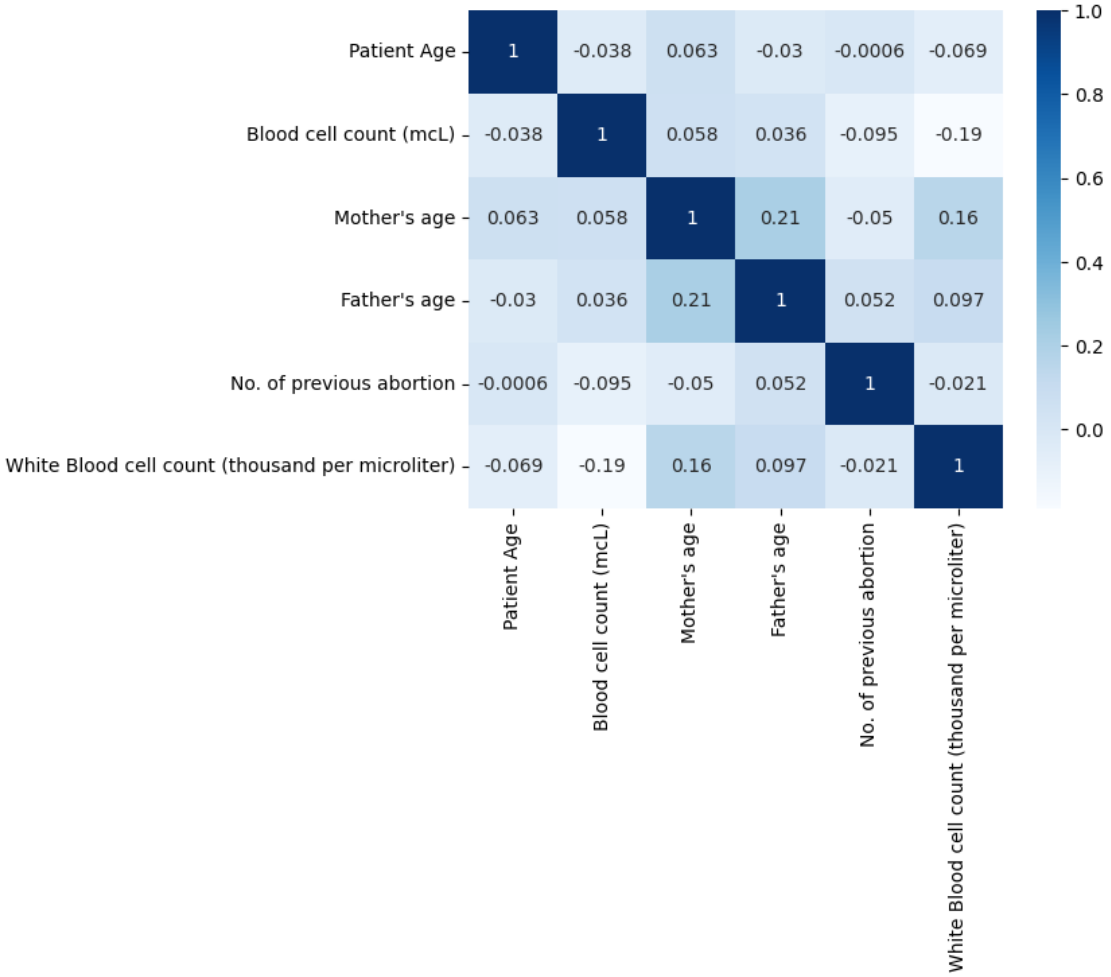
[62]:

	Patient Age	Blood cell count (mcL)	Mother's age	Father's age	No. of previous abortion
10376	14.0	4.750030	28.000000	40.0	3.000000
6353	6.0	4.567651	35.000000	61.0	2.000000
11223	14.0	4.911487	48.000000	29.0	2.003062
18956	8.0	4.633470	34.526454	60.0	
6256	6.0	4.744146	19.000000	59.0	

```
18956                    3.000000
6256                     4.000000

       White Blood cell count (thousand per microliter)
10376                                          9.313145
6353                                           9.650795
11223                                          6.012206
18956                                          9.129951
6256                                           5.288666
```

[63]:
```
corr_matrix = num2.corr()
sns.heatmap(corr_matrix, annot=True, cmap='Blues')
```

[63]: <Axes: >

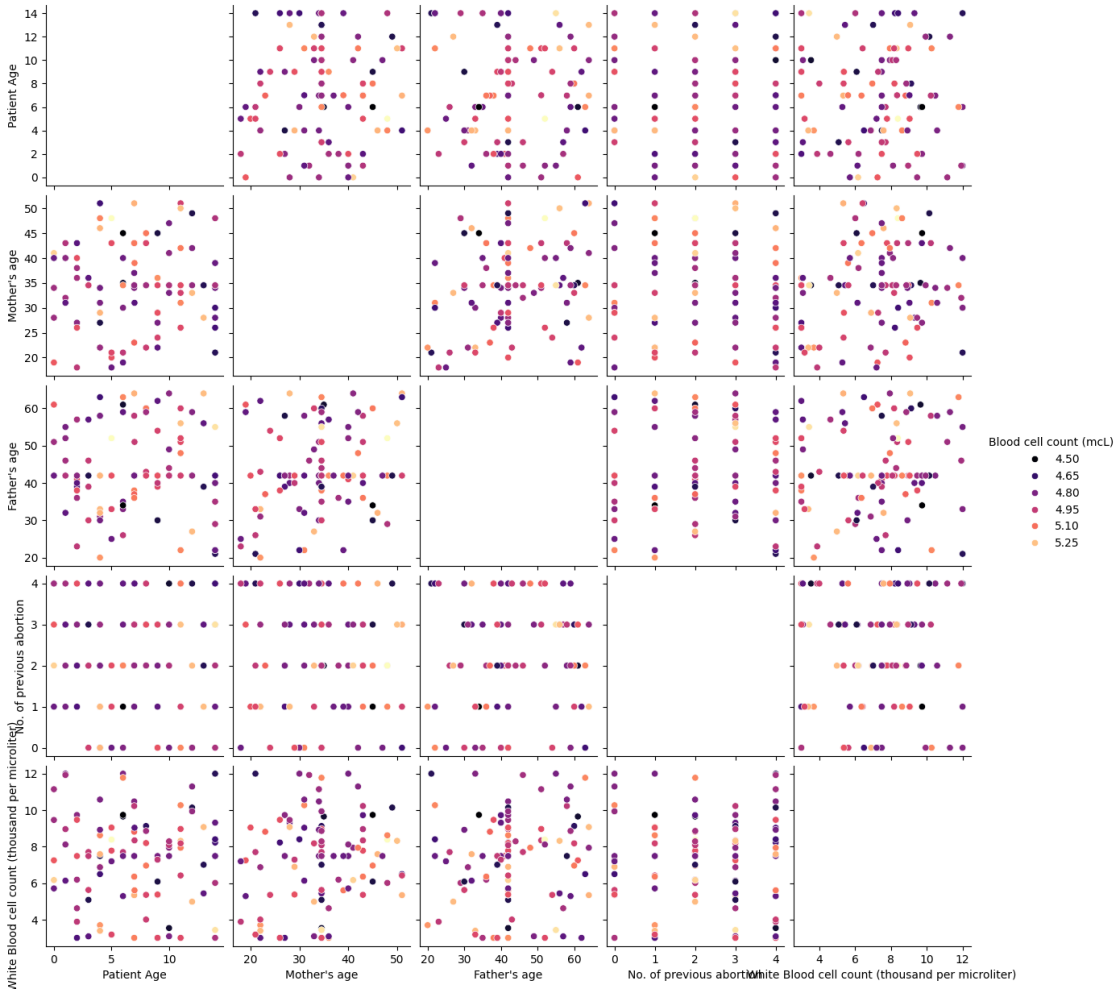


[64]:
```
num.columns
```

[64]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)', 'Symptom 1',
       'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5'],
      dtype='object')
```

[65]:
```
sns.pairplot(num2, diag_kind ='kde',hue ='Blood cell count (mcL)', ␣
↪palette='magma')
```

[65]:
```
<seaborn.axisgrid.PairGrid at 0x7e473485c710>
```



[66]:
```
num2.columns
```

[66]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)'],
```

```
      dtype='object')
```

[67]:
```
cat.columns
```

[67]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

[68]:
```
gb = df.groupby('Genetic Disorder')['Birth defects'].value_counts()
gb
```

[68]:
```
Genetic Disorder                                Birth defects
Mitochondrial genetic inheritance disorders     Multiple         4666
                                                Singular         4612
Multifactorial genetic inheritance disorders    Singular          978
                                                Multiple          907
Single-gene inheritance diseases                Multiple         3483
                                                Singular         3452
Name: count, dtype: int64
```

[69]:
```
tests = df[['Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5']]
tests.head()
```

[69]:

	Test 1	Test 2	Test 3	Test 4	Test 5
0	0.0	NaN	NaN	1.0	0.0
1	NaN	0.0	0.0	1.0	0.0
2	0.0	0.0	0.0	1.0	0.0
3	0.0	0.0	0.0	1.0	0.0
4	0.0	0.0	0.0	1.0	0.0

[70]:
```
tests.fillna(tests.mode(), inplace=True)
```

[71]:
```
tests.head()
```

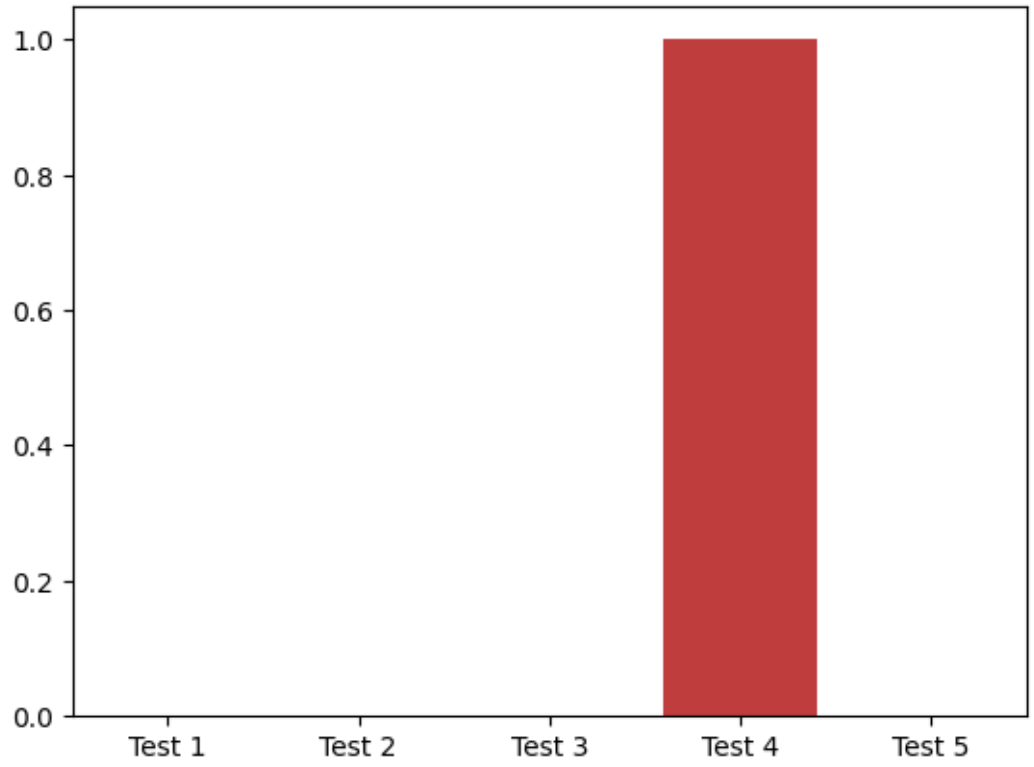
[71]:

	Test 1	Test 2	Test 3	Test 4	Test 5
0	0.0	0.0	0.0	1.0	0.0
1	NaN	0.0	0.0	1.0	0.0
2	0.0	0.0	0.0	1.0	0.0
3	0.0	0.0	0.0	1.0	0.0

```
4     0.0     0.0     0.0     1.0     0.0
```

[72]:
```
sns.barplot(tests)
```

[72]:
```
<Axes: >
```



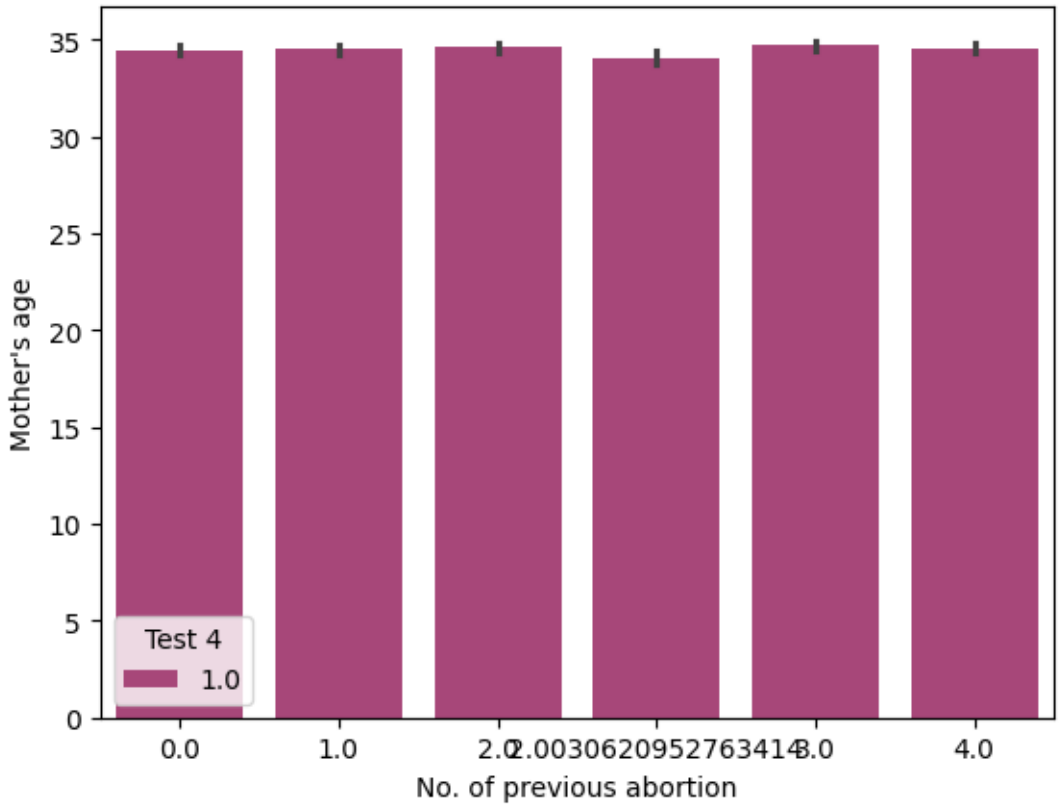
[73]:
```
tests.value_counts()
tests = num['Test 4']
```

[74]:
```
num.columns
```

[74]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)', 'Symptom 1',
       'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5'],
      dtype='object')
```

[75]:
```
sns.barplot(x = num['No. of previous abortion'], y=num["Mother's age"], hue =␣
 ↪tests, palette='magma' )
```

[75]:
```
<Axes: xlabel='No. of previous abortion', ylabel="Mother's age">
```



[76]:
```
num['No. of previous abortion'].value_counts()
```

[76]:
```
No. of previous abortion
2.000000    4117
4.000000    4005
0.000000    3964
1.000000    3928
3.000000    3907
2.003062    2162
Name: count, dtype: int64
```

[77]:
```
df_abortion = num['No. of previous abortion']
df_abortion.head()
```

[77]:
```
0    2.003062
1    2.003062
2    4.000000
3    1.000000
4    4.000000
Name: No. of previous abortion, dtype: float64
```

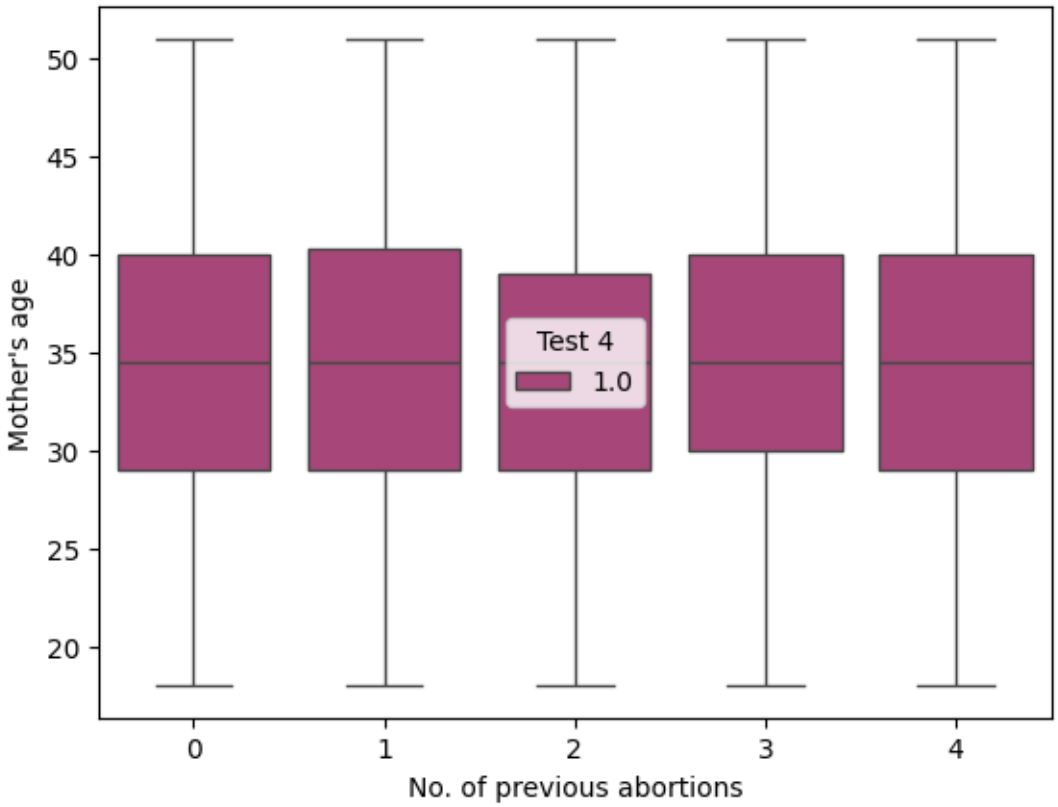
[78]:
```
df_abortion_int = df_abortion.astype(int)
df_abortion_int.value_counts()
```

[78]:
```
No. of previous abortion
2    6279
4    4005
0    3964
1    3928
3    3907
Name: count, dtype: int64
```

[79]:
```
num['No. of previous abortions'] = df_abortion_int
```

[80]:
```
sns.boxplot(x = num['No. of previous abortions'], y=num["Mother's age"], hue =␣
↪tests, palette='magma' )
```

[80]:
```
<Axes: xlabel='No. of previous abortions', ylabel="Mother's age">
```



[81]:
```
num.columns
```

[81]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)', 'Symptom 1',
       'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'No. of previous abortions'],
      dtype='object')
```

[82]:
```
symptoms = df[['Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4',
       'Symptom 5']]
symptoms.head()
```

[82]:

	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5
0	1.0	1.0	1.0	1.0	1.0
1	1.0	NaN	1.0	1.0	0.0
2	0.0	1.0	1.0	1.0	1.0
3	0.0	0.0	1.0	0.0	0.0
4	0.0	0.0	0.0	0.0	NaN

[83]:
```
corr_symptoms = symptoms.corr()
corr_symptoms.style.background_gradient(cmap='magma')
```

[83]:
```
<pandas.io.formats.style.Styler at 0x7e47362bdf50>
```

DISEASE ANALYSIS

[84]:
```
df.head()
```

[84]:

	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Mother's age	Father's age	Institute Name
0	2.0	Yes	No	Yes	No	4.760603	NaN	NaN	Boston Specialty & Rehabilitation Hospital
1	4.0	Yes	Yes	No	No	4.910669	NaN	23.0	St. Margaret's Hospital For Women
2	6.0	Yes	No	No	No	4.893297	41.0	22.0	NaN
3	12.0	Yes	No	Yes	No	4.705280	21.0	NaN	NaN
4	11.0	Yes	No	NaN	Yes	4.720703	32.0	NaN	Carney Hospital

```
                              Location of Institute  … Birth defects  \
0  55 FRUIT ST\nCENTRAL, MA 02114\n(42.3624748574…  …           NaN
1  1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 021…  …      Multiple
2                                                -  …      Singular
3  55 FRUIT ST\nCENTRAL, MA 02114\n(42.3624748574…  …      Singular
4  300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42…  …      Multiple

  White Blood cell count (thousand per microliter) Blood test result  \
0                                        9.857562               NaN
1                                        5.522560            normal
2                                             NaN            normal
3                                        7.919321      inconclusive
4                                        4.098210               NaN

   Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
0        1.0        1.0        1.0        1.0        1.0
1        1.0        NaN        1.0        1.0        0.0
2        0.0        1.0        1.0        1.0        1.0
3        0.0        0.0        1.0        0.0        0.0
4        0.0        0.0        0.0        0.0        NaN

                              Genetic Disorder  \
0   Mitochondrial genetic inheritance disorders
1                                           NaN
2  Multifactorial genetic inheritance disorders
3   Mitochondrial genetic inheritance disorders
4  Multifactorial genetic inheritance disorders

                    Disorder Subclass
0  Leber's hereditary optic neuropathy
1                     Cystic fibrosis
2                            Diabetes
3                      Leigh syndrome
4                              Cancer

[5 rows x 41 columns]
```

[85]:
```
df_disorders = df.groupby('Disorder Subclass')['Genetic Disorder'].
↪value_counts()
df_disorders
```

[85]:
```
Disorder Subclass                  Genetic Disorder
Alzheimer's                        Multifactorial genetic inheritance
disorders     133
Cancer                             Multifactorial genetic inheritance
disorders      91
```

```
Cystic fibrosis                          Single-gene inheritance diseases
3145
Diabetes                                 Multifactorial genetic inheritance
disorders    1653
Hemochromatosis                          Single-gene inheritance diseases
1228
Leber's hereditary optic neuropathy      Mitochondrial genetic inheritance disorders
587
Leigh syndrome                           Mitochondrial genetic inheritance disorders
4683
Mitochondrial myopathy                   Mitochondrial genetic inheritance disorders
3971
Tay-Sachs                                Single-gene inheritance diseases
2556
Name: count, dtype: int64
```

[86]:
```
cancer = df[df['Disorder Subclass'] == 'Cancer']
cancer.head()
```

[86]:
```
     Patient Age Genes in mother's side Inherited from father Maternal gene  \
4           11.0                    Yes                    No           NaN
107         13.0                    Yes                    No            No
283         12.0                    Yes                    No            No
304          4.0                     No                    No            No
513          2.0                    Yes                    No           Yes

    Paternal gene  Blood cell count (mcL)  Mother's age  Father's age  \
4             Yes                4.720703          32.0           NaN
107            No                4.970532          42.0          24.0
283            No                5.015183           NaN           NaN
304            No                4.688240          45.0           NaN
513            No                5.058617          48.0          57.0

                Institute Name  \
4              Carney Hospital
107  Lemuel Shattuck Hospital
283            Not applicable
304                       NaN
513            Not applicable

                                Location of Institute  … Birth defects  \
4    300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42…  …      Multiple
107  1153 CENTRE ST\nJAMAICA PLAIN, MA 02130\n(42.3…  …      Singular
283                                                -  …           NaN
304  300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42…  …           NaN
513                                                -  …      Singular
```

```
    White Blood cell count (thousand per microliter) Blood test result  \
4                                          4.098210               NaN
107                                       10.711108            normal
283                                        3.480522          abnormal
304                                       11.531087      inconclusive
513                                        6.327895            normal

     Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
4          0.0        0.0        0.0        0.0        NaN
107        0.0        0.0        0.0        0.0        0.0
283        NaN        0.0        NaN        0.0        0.0
304        0.0        NaN        0.0        0.0        0.0
513        0.0        0.0        NaN        0.0        0.0

                                Genetic Disorder Disorder Subclass
4    Multifactorial genetic inheritance disorders            Cancer
107  Multifactorial genetic inheritance disorders            Cancer
283  Multifactorial genetic inheritance disorders            Cancer
304  Multifactorial genetic inheritance disorders            Cancer
513  Multifactorial genetic inheritance disorders            Cancer

[5 rows x 41 columns]
```

[87]:
```
cancer_num = cancer.select_dtypes(exclude='object')
cancer_num.shape
```

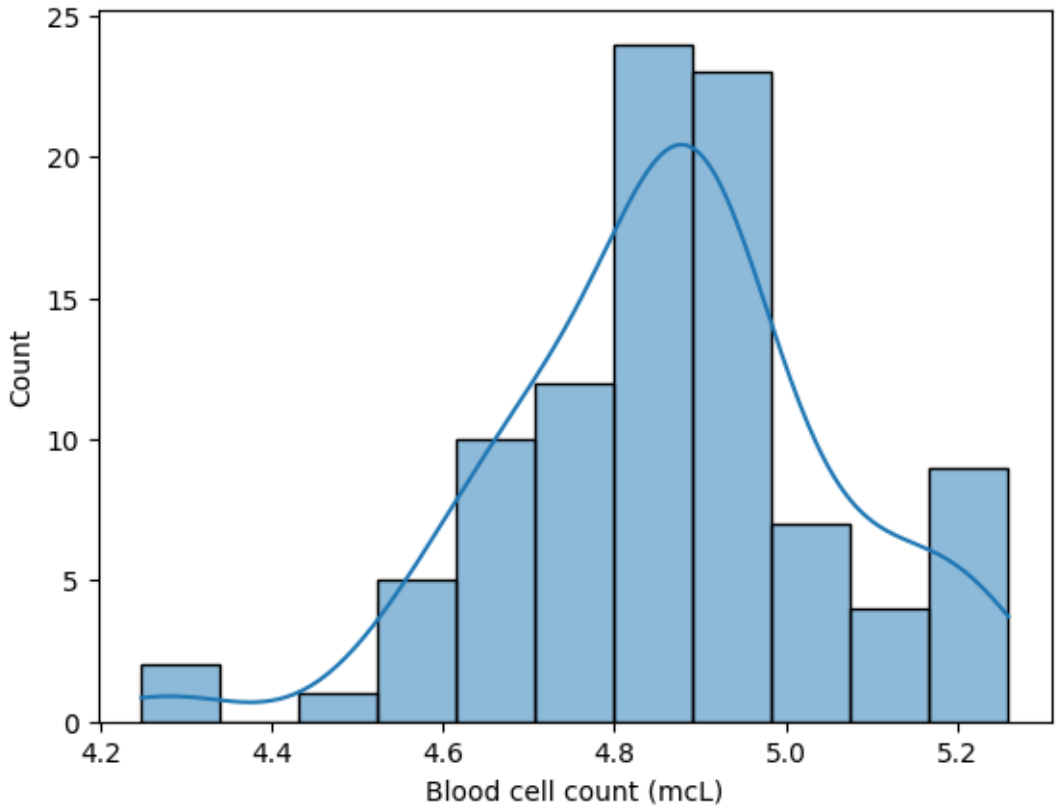
[87]:
```
(97, 16)
```

[88]:
```
cancer_num.columns
```

[88]:
```
Index(['Patient Age', 'Blood cell count (mcL)', 'Mother's age', 'Father's age',
       'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5',
       'No. of previous abortion',
       'White Blood cell count (thousand per microliter)', 'Symptom 1',
       'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5'],
      dtype='object')
```

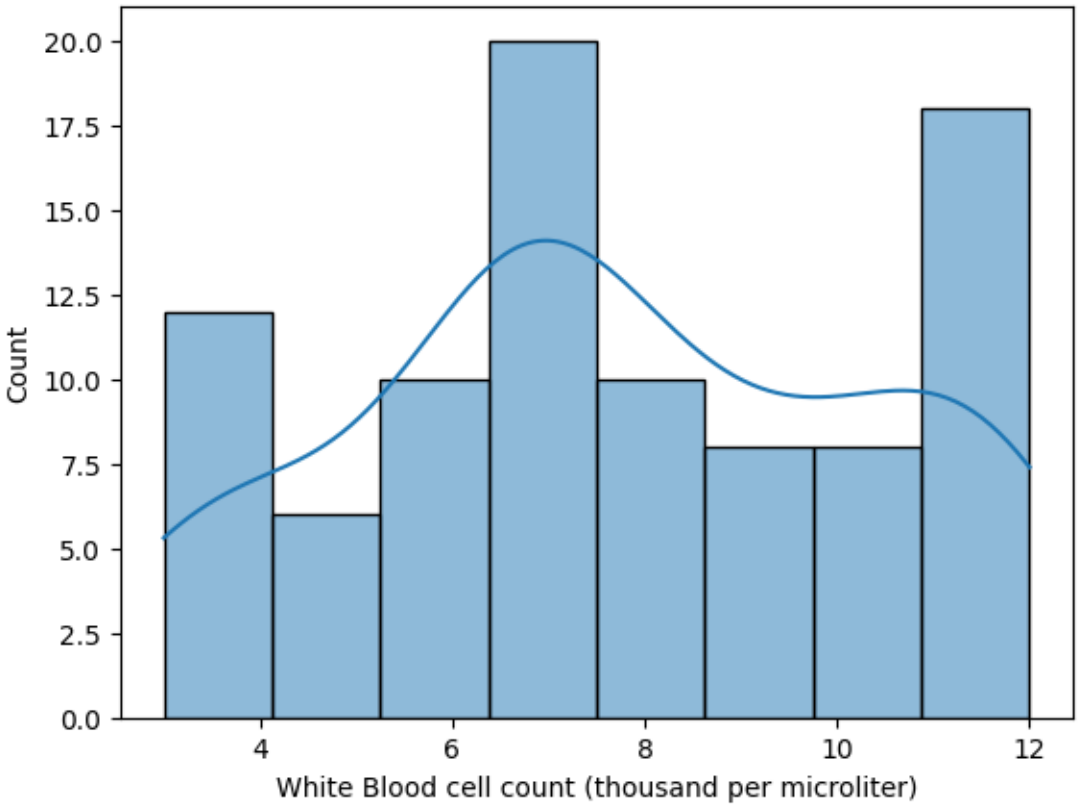
[89]:
```
sns.histplot(x='Blood cell count (mcL)', kde=True, data=cancer_num,␣
↪palette='magma')
```

[89]:
```
<Axes: xlabel='Blood cell count (mcL)', ylabel='Count'>
```



[90]:
```
sns.histplot(x='White Blood cell count (thousand per microliter)', kde=True,␣
↪data=cancer_num, palette='magma')
```

[90]:
```
<Axes: xlabel='White Blood cell count (thousand per microliter)',
ylabel='Count'>
```



[91]:
```
# cancer and maternal and paternal genes

cancer_genes = cancer[['Maternal gene', 'Paternal gene']]
cancer_genes
```

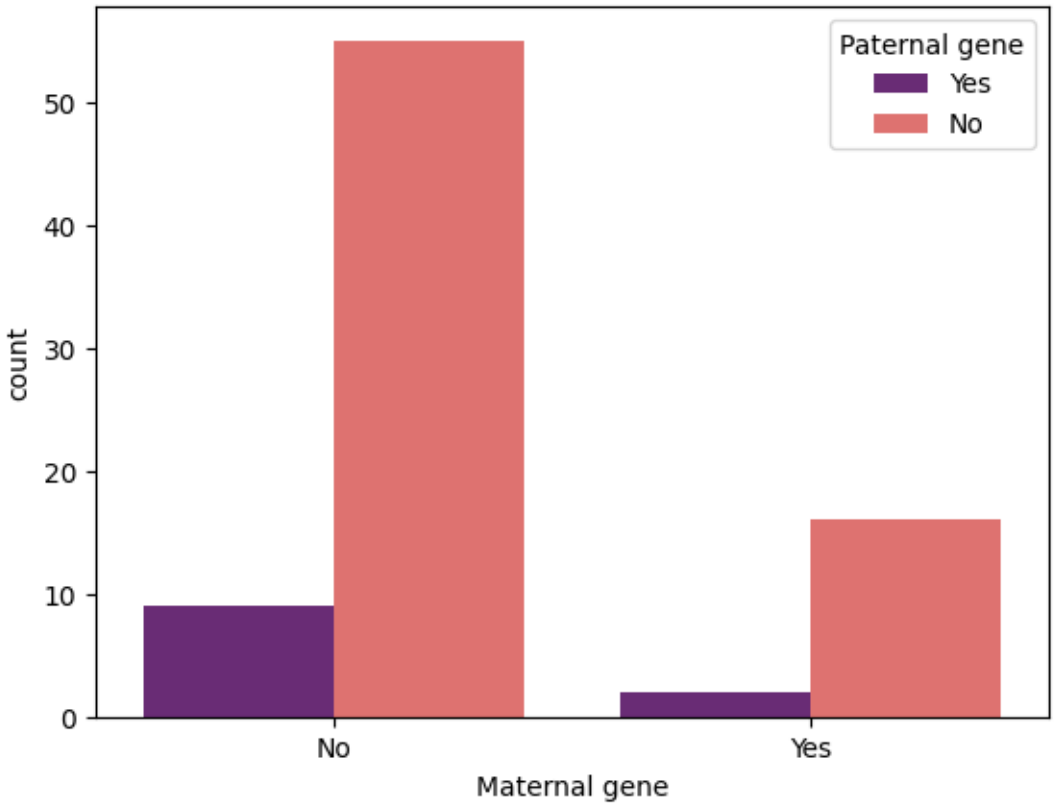
[91]:

	Maternal gene	Paternal gene
4	NaN	Yes
107	No	No
283	No	No
304	No	No
513	Yes	No
…	…	…
21185	No	No
21356	No	No
21408	No	No
21478	No	No
21944	Yes	No

[97 rows x 2 columns]

[92]:
```
sns.countplot(x='Maternal gene',hue='Paternal gene', data=cancer_genes,␣
↪palette='magma')
```

[92]: <Axes: xlabel='Maternal gene', ylabel='count'>



[93]:
```
# symptoms and cancer

cancer_symptoms = cancer[['Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4',
       'Symptom 5']]
cancer_symptoms.value_counts()

cancer_symptoms = cancer_symptoms.dropna()
cancer_symptoms.head()
```

[93]:

	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5
107	0.0	0.0	0.0	0.0	0.0
553	1.0	0.0	0.0	0.0	0.0
1162	0.0	0.0	0.0	0.0	0.0
2139	0.0	0.0	0.0	0.0	0.0
2267	0.0	0.0	1.0	0.0	0.0

[94]:
```
df_disorders
```

[94]:
```
Disorder Subclass                    Genetic Disorder
Alzheimer's                          Multifactorial genetic inheritance
disorders     133
Cancer                               Multifactorial genetic inheritance
disorders      91
Cystic fibrosis                      Single-gene inheritance diseases
3145
Diabetes                             Multifactorial genetic inheritance
disorders    1653
Hemochromatosis                      Single-gene inheritance diseases
1228
Leber's hereditary optic neuropathy  Mitochondrial genetic inheritance disorders
587
Leigh syndrome                       Mitochondrial genetic inheritance disorders
4683
Mitochondrial myopathy               Mitochondrial genetic inheritance disorders
3971
Tay-Sachs                            Single-gene inheritance diseases
2556
Name: count, dtype: int64
```

[95]:
```
df_disorders["Alzheimer's"]
```

[95]:
```
Genetic Disorder
Multifactorial genetic inheritance disorders    133
Name: count, dtype: int64
```

[96]:
```
alzheimer = df[df['Disorder Subclass'] == "Alzheimer's"]
alzheimer.head()
```

[96]:
```
     Patient Age Genes in mother's side Inherited from father Maternal gene  \
202         14.0                   Yes                    Yes           Yes
306          9.0                   Yes                    Yes           NaN
380         10.0                   Yes                     No            No
405          0.0                   Yes                    Yes           Yes
525          6.0                   Yes                    Yes            No

    Paternal gene  Blood cell count (mcL)  Mother's age  Father's age  \
202           Yes                4.826227          47.0          58.0
306           Yes                4.870173           NaN           NaN
380           Yes                4.962701          19.0           NaN
405           Yes                4.687219          29.0          44.0
525            No                4.704889          34.0          57.0

                        Institute Name  \
```

```
202  Massachusetts Eye & Ear Infirmary
306               Boston City Hospital
380                        Va Hospital
405                     Not applicable
525           Shriners Burns Institute

                                Location of Institute  … Birth defects  \
202  1200 Centre St\nRoslindale, MA 02131\n(42.2973…  …      Multiple
306  750 WASHINGTON ST\nCENTRAL, MA 02111\n(42.3499…  …           NaN
380  300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42…  …      Singular
405                                                -  …           NaN
525  249 RIVER ST\nMATTAPAN, MA 02126\n(42.27137912…  …      Singular

    White Blood cell count (thousand per microliter)  Blood test result  \
202                                         5.178027       inconclusive
306                                         5.497112  slightly abnormal
380                                         3.000000  slightly abnormal
405                                        10.270923       inconclusive
525                                         9.024526  slightly abnormal

     Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
202        1.0        1.0        1.0        1.0        1.0
306        1.0        1.0        NaN        1.0        1.0
380        0.0        1.0        1.0        1.0        1.0
405        1.0        1.0        1.0        1.0        1.0
525        1.0        1.0        1.0        1.0        NaN

                                Genetic Disorder Disorder Subclass
202  Multifactorial genetic inheritance disorders       Alzheimer's
306  Multifactorial genetic inheritance disorders       Alzheimer's
380  Multifactorial genetic inheritance disorders       Alzheimer's
405  Multifactorial genetic inheritance disorders       Alzheimer's
525  Multifactorial genetic inheritance disorders       Alzheimer's

[5 rows x 41 columns]
```

[97]:
```
# birth defects vs alzheimers

alzheimer_birth = alzheimer['Birth defects']
alzheimer_birth.value_counts()
```

[97]:
```
Birth defects
Singular    81
Multiple    59
Name: count, dtype: int64
```

[98]:
```
df['Institute Name'].value_counts()
```

[98]:
```
Institute Name
Not applicable                                    8440
Franciscan Children's  Hospital                    363
Carney Hospital                                    357
New England Medical Center                         350
Hebrew Rehabilitation Center                       349
VA Hospital                                        344
Shriners Burns Institute                           341
Massachusetts Eye & Ear Infirmary                  337
Brigham And Women's Hospital                       334
Boston City Hospital                               330
St. Margaret's Hospital For Women                  329
Arbour Hospital                                    327
Spaulding Rehabilitation Hospital                  325
Faulkner Hospital                                  325
Children's Hospital                                324
Kindred Hospital                                   324
Dana-farber Cancer Institute                       323
Boston Specialty & Rehabilitation Hospital         322
Massachusetts General Hospital                     321
Beth Israel Deaconess Medical Center East Cam      320
Boston Medical Center                              318
New England Baptist Hospital                       317
Beth Israel Deaconess Medical Center West Cam      315
Jewish Memorial Hospital                           315
Lemuel Shattuck Hospital                           313
Va Hospital                                        312
St. Elizabeth's Hospital                           302
Name: count, dtype: int64
```

[99]:
```
mass_general = df[df['Institute Name'] == "Massachusetts General Hospital"]
mass_general
```

[99]:
```
       Patient Age Genes in mother's side Inherited from father Maternal gene  \
5             14.0                   Yes                    No           Yes
93             9.0                    No                    No           Yes
166            2.0                   Yes                    No           Yes
232           14.0                    No                    No            No
235            0.0                    No                    No            No
…              …                      …                     …             …
21827          8.0                   Yes                   Yes            No
21832         12.0                   Yes                   Yes           Yes
21852          NaN                    No                    No            No
21920          7.0                   Yes                   Yes           Yes
22064          2.0                   Yes                   Yes           NaN

      Paternal gene  Blood cell count (mcL)  Mother's age  Father's age  \
```

```
5                No                 5.103188          NaN          NaN
93               No                 4.999788         51.0         27.0
166             Yes                 4.988560         47.0         49.0
232              No                 4.987643         40.0         27.0
235             Yes                 5.103858          NaN         39.0
…               …                    …             …           …
21827            No                 4.747883         24.0         46.0
21832           Yes                 4.840895         46.0         30.0
21852            No                 4.845462         19.0          NaN
21920            No                 4.994396         42.0         31.0
22064            No                 5.168511          NaN          NaN

                      Institute Name  \
5      Massachusetts General Hospital
93     Massachusetts General Hospital
166    Massachusetts General Hospital
232    Massachusetts General Hospital
235    Massachusetts General Hospital
…                                 …
21827  Massachusetts General Hospital
21832  Massachusetts General Hospital
21852  Massachusetts General Hospital
21920  Massachusetts General Hospital
22064  Massachusetts General Hospital

                                    Location of Institute  … Birth defects  \
5      55 FRUIT ST\nCENTRAL, MA 02114\n(42.3624748574…  …      Multiple
93     750 WASHINGTON ST\nCENTRAL, MA 02111\n(42.3499…  …      Multiple
166    1400 VFW Parkway\nWest Roxbury, MA 02132\n(42…  …      Singular
232    1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 021…  …      Multiple
235    1400 VFW Parkway\nWest Roxbury, MA 02132\n(42…  …      Singular
…                                                    …  …          …
21827  818 HARRISON AV\nSOUTH END, MA 02118\n(42.3359…  …      Multiple
21832  125 PARKER HILL AV\nJAMAICA PLAIN, MA 02120\n(…  …      Multiple
21852  75 FRANCIS ST\nFENWAY/KENMORE, MA 02115\n(42.3…  …      Multiple
21920  75 FRANCIS ST\nFENWAY/KENMORE, MA 02115\n(42.3…  …      Singular
22064  818 HARRISON AV\nSOUTH END, MA 02118\n(42.3359…  …      Multiple

       White Blood cell count (thousand per microliter)  Blood test result  \
5                                              10.272230             normal
93                                              6.746098             normal
166                                             5.467502           abnormal
232                                            10.288447             normal
235                                             3.585694                NaN
…                                                  …                …
21827                                          11.849304             normal
21832                                           8.331476  slightly abnormal
```

```
21852                                        8.949926       inconclusive
21920                                        3.000000           abnormal
22064                                        9.826694  slightly abnormal

       Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
5            1.0        0.0        0.0        1.0        0.0
93           1.0        1.0        1.0        0.0        1.0
166          1.0        0.0        1.0        0.0        0.0
232          1.0        1.0        1.0        1.0        0.0
235          1.0        1.0        1.0        0.0        0.0
…            …          …          …          …          …
21827        0.0        1.0        0.0        1.0        1.0
21832        0.0        0.0        1.0        0.0        1.0
21852        0.0        NaN        0.0        NaN        1.0
21920        1.0        1.0        0.0        1.0        1.0
22064        1.0        0.0        1.0        0.0        0.0

                                Genetic Disorder  \
5                 Single-gene inheritance diseases
93     Mitochondrial genetic inheritance disorders
166    Mitochondrial genetic inheritance disorders
232    Mitochondrial genetic inheritance disorders
235    Mitochondrial genetic inheritance disorders
…                                              …
21827             Single-gene inheritance diseases
21832  Mitochondrial genetic inheritance disorders
21852  Mitochondrial genetic inheritance disorders
21920  Mitochondrial genetic inheritance disorders
22064  Mitochondrial genetic inheritance disorders

                         Disorder Subclass
5                          Cystic fibrosis
93                          Leigh syndrome
166                         Leigh syndrome
232                                    NaN
235                         Leigh syndrome
…                                       …
21827                      Cystic fibrosis
21832                       Leigh syndrome
21852                Mitochondrial myopathy
21920  Leber's hereditary optic neuropathy
22064                       Leigh syndrome

[321 rows x 41 columns]
```

[100]:
```
df_disorders
```

[100]:
```
Disorder Subclass                    Genetic Disorder
Alzheimer's                          Multifactorial genetic inheritance
disorders     133
Cancer                               Multifactorial genetic inheritance
disorders      91
Cystic fibrosis                      Single-gene inheritance diseases
3145
Diabetes                             Multifactorial genetic inheritance
disorders    1653
Hemochromatosis                      Single-gene inheritance diseases
1228
Leber's hereditary optic neuropathy  Mitochondrial genetic inheritance disorders
587
Leigh syndrome                       Mitochondrial genetic inheritance disorders
4683
Mitochondrial myopathy               Mitochondrial genetic inheritance disorders
3971
Tay-Sachs                            Single-gene inheritance diseases
2556
Name: count, dtype: int64
```

[101]:
```
#cystic fibrosis

cf = df[df['Disorder Subclass'] == 'Cystic fibrosis']
cf.head()
```

[101]:
```
    Patient Age Genes in mother's side Inherited from father Maternal gene  \
1           4.0                    Yes                    Yes            No
5          14.0                    Yes                     No           Yes
11          7.0                     No                     No            No
16          0.0                    Yes                    Yes            No
17          NaN                     No                     No            No

   Paternal gene  Blood cell count (mcL)  Mother's age  Father's age  \
1             No                4.910669           NaN          23.0
5             No                5.103188           NaN           NaN
11           Yes                4.848795           NaN           NaN
16            No                4.798520           NaN          57.0
17            No                4.952457         24.0           24.0

                      Institute Name  \
1   St. Margaret's Hospital For Women
5      Massachusetts General Hospital
11                    Not applicable
16         New England Medical Center
17                               NaN
```

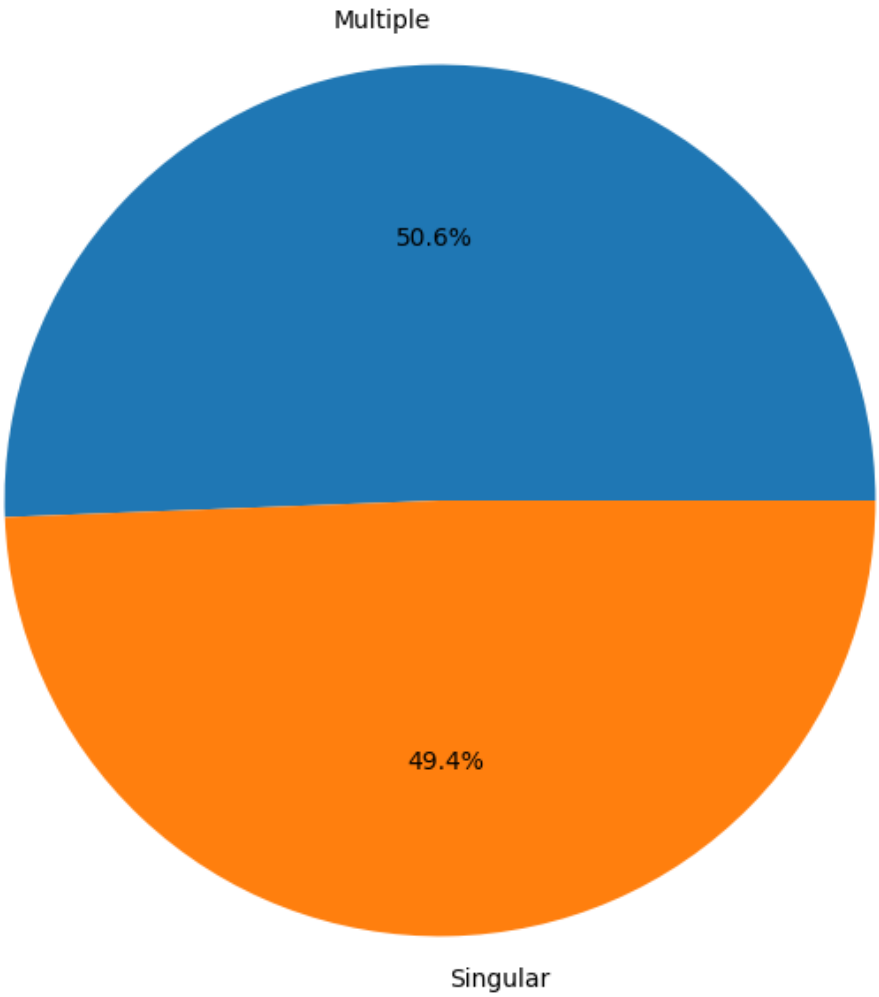
```
                                Location of Institute  … Birth defects  \
1   1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 021…  …      Multiple
5   55 FRUIT ST\nCENTRAL, MA 02114\n(42.3624748574…  …      Multiple
11                                                 -  …      Multiple
16  125 PARKER HILL AV\nJAMAICA PLAIN, MA 02120\n(…  …      Multiple
17                                                 -  …           NaN

   White Blood cell count (thousand per microliter)  Blood test result  \
1                                          5.522560             normal
5                                         10.272230             normal
11                                         8.409691  slightly abnormal
16                                              NaN             normal
17                                        10.031078       inconclusive

    Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
1         1.0        NaN        1.0        1.0        0.0
5         1.0        0.0        0.0        1.0        0.0
11        0.0        1.0        1.0        1.0        1.0
16        1.0        1.0        1.0        NaN        1.0
17        1.0        1.0        0.0        1.0        1.0

                   Genetic Disorder Disorder Subclass
1                               NaN   Cystic fibrosis
5   Single-gene inheritance diseases   Cystic fibrosis
11  Single-gene inheritance diseases   Cystic fibrosis
16  Single-gene inheritance diseases   Cystic fibrosis
17  Single-gene inheritance diseases   Cystic fibrosis

[5 rows x 41 columns]
```

[102]:
```
cf['Birth defects'].value_counts()
```

[102]:
```
Birth defects
Multiple    1581
Singular    1543
Name: count, dtype: int64
```

[103]:
```
plt.figure(figsize=(10,8))
plt.pie(cf['Birth defects'].value_counts(), labels=cf['Birth defects'].
 ↪value_counts().index, autopct='%1.1f%%')
```

[103]:
```
([<matplotlib.patches.Wedge at 0x7e472ad527d0>,
  <matplotlib.patches.Wedge at 0x7e472ad95a10>],
 [Text(-0.021016392664593302, 1.0997992140565331, 'Multiple'),
  Text(0.021016392664593167, -1.0997992140565331, 'Singular')],
 [Text(-0.011463486907959982, 0.5998904803944726, '50.6%'),
  Text(0.011463486907959907, -0.5998904803944726, '49.4%')])
```



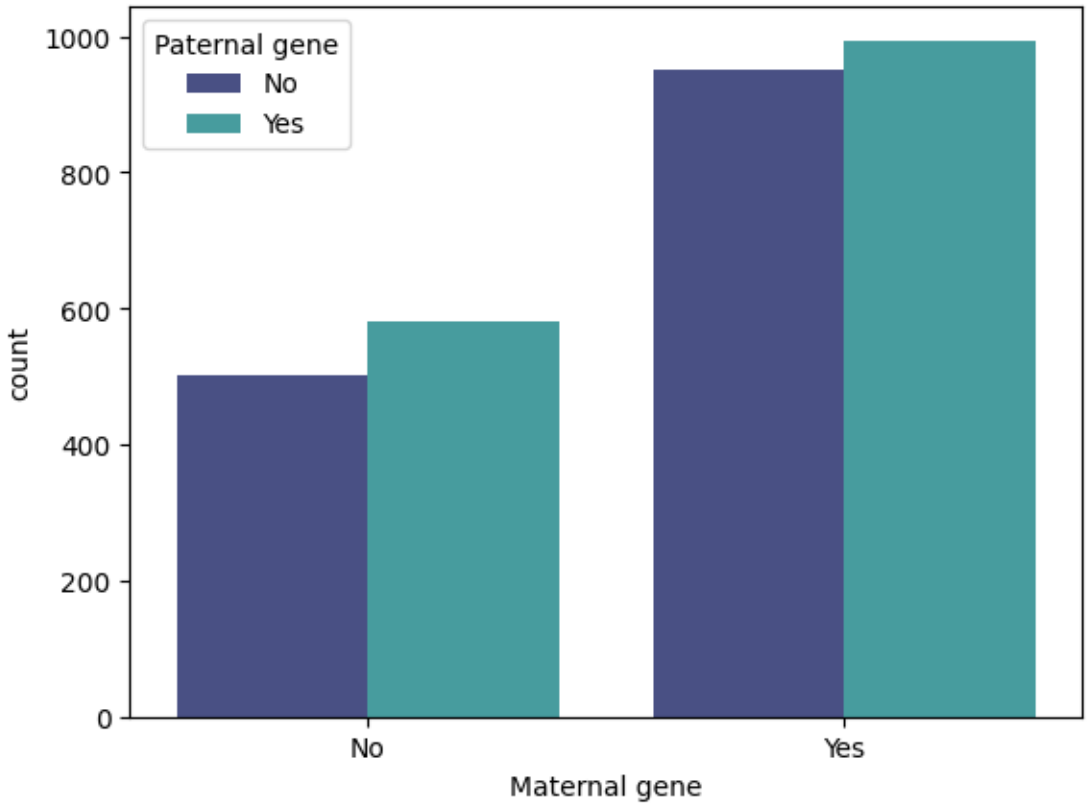
[104]:
```
# cf and genes

cf_genes = cf[['Maternal gene', 'Paternal gene']]
cf_genes.head()
```

[104]:

	Maternal gene	Paternal gene
1	No	No
5	Yes	No
11	No	Yes
16	No	No
17	No	No

[105]:
```
sns.countplot(x='Maternal gene',hue='Paternal gene', data=cf_genes,␣
↪palette='mako')
```

[105]: <Axes: xlabel='Maternal gene', ylabel='count'>



[106]:
```
df.columns
```

[106]:
```
Index(['Patient Age', 'Genes in mother's side', 'Inherited from father',
       'Maternal gene', 'Paternal gene', 'Blood cell count (mcL)',
       'Mother's age', 'Father's age', 'Institute Name',
       'Location of Institute', 'Status', 'Respiratory Rate (breaths/min)',
       'Heart Rate (rates/min', 'Test 1', 'Test 2', 'Test 3', 'Test 4',
       'Test 5', 'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
```

```
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

[107]:
```
cat.columns
```

[107]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Location of Institute', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min',
       'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

[111]:
```
encoded = pd.get_dummies(cat, columns=['Respiratory Rate (breaths/min)', 'Heart␣
 ↪Rate (rates/min']) # Corrected column name
encoded.head()
```

[111]:

	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	\
0	Yes	No	Yes	No	
1	Yes	Yes	No	No	
2	Yes	No	No	No	
3	Yes	No	Yes	No	
4	Yes	No	Yes	Yes	

	Institute Name	\
0	Boston Specialty & Rehabilitation Hospital	
1	St. Margaret's Hospital For Women	
2	Not applicable	
3	Not applicable	
4	Carney Hospital	

	Location of Institute	Status	\
0	55 FRUIT ST\nCENTRAL, MA 02114\n(42.3624748574…	Alive	
1	1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 021…	Deceased	
2	-	Alive	
3	55 FRUIT ST\nCENTRAL, MA 02114\n(42.3624748574…	Deceased	
4	300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42…	Alive	

	Parental consent	Follow-up	Gender	…	Assisted conception IVF/ART	\
0	Yes	High	Ambiguous	…	No	
1	Yes	High	Ambiguous	…	No	
2	Yes	Low	Ambiguous	…	Yes	
3	Yes	High	Male	…	Yes	

```
4              Yes       Low       Male  …                          Yes

  History of anomalies in previous pregnancies Birth defects  \
0                                          Yes      Singular
1                                          Yes      Multiple
2                                          Yes      Singular
3                                          Yes      Singular
4                                           No      Multiple

   Blood test result                                Genetic Disorder  \
0  slightly abnormal   Mitochondrial genetic inheritance disorders
1             normal   Mitochondrial genetic inheritance disorders
2             normal  Multifactorial genetic inheritance disorders
3       inconclusive   Mitochondrial genetic inheritance disorders
4  slightly abnormal  Multifactorial genetic inheritance disorders

                    Disorder Subclass  \
0  Leber's hereditary optic neuropathy
1                      Cystic fibrosis
2                             Diabetes
3                       Leigh syndrome
4                               Cancer

  Respiratory Rate (breaths/min)_Normal (30-60)  \
0                                          True
1                                         False
2                                          True
3                                         False
4                                         False

  Respiratory Rate (breaths/min)_Tachypnea Heart Rate (rates/min_Normal  \
0                                   False                          True
1                                    True                          True
2                                   False                         False
3                                    True                          True
4                                    True                         False

  Heart Rate (rates/min_Tachycardia
0                             False
1                             False
2                              True
3                             False
4                              True

[5 rows x 27 columns]
```

[112]:
```
encoded = encoded.drop(columns='Location of Institute', axis=1)
```

[113]:
```
encoded.columns
```

[113]:
```
Index(['Genes in mother's side', 'Inherited from father', 'Maternal gene',
       'Paternal gene', 'Institute Name', 'Status', 'Parental consent',
       'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies', 'Birth defects',
       'Blood test result', 'Genetic Disorder', 'Disorder Subclass',
       'Respiratory Rate (breaths/min)_Normal (30-60)',
       'Respiratory Rate (breaths/min)_Tachypnea',
       'Heart Rate (rates/min_Normal', 'Heart Rate (rates/min_Tachycardia'],
      dtype='object')
```

[114]:
```
mental = df['H/O serious maternal illness']
mental.head()
```

[114]:
```
0    NaN
1    Yes
2     No
3    Yes
4    Yes
Name: H/O serious maternal illness, dtype: object
```

[115]:
```
mental_illness = mental.dropna()
mental_illness.head()
```

[115]:
```
1    Yes
2     No
3    Yes
4    Yes
5     No
Name: H/O serious maternal illness, dtype: object
```

[116]:
```
mental_illness.value_counts()
```

[116]:
```
H/O serious maternal illness
No     10012
Yes     9919
Name: count, dtype: int64
```

[119]:
```
cancer_and_mental_illness = df[df['H/O serious maternal illness'].notna() &␣
 ↪(df['Disorder Subclass'] == 'Cancer')]
```

[120]:
```
cancer_and_mental_illness.head()
```

[120]:
```
     Patient Age Genes in mother's side Inherited from father Maternal gene  \
4           11.0                    Yes                     No           NaN
283         12.0                    Yes                     No            No
304          4.0                     No                     No            No
513          2.0                    Yes                     No           Yes
553          7.0                    Yes                     No           NaN

    Paternal gene  Blood cell count (mcL)  Mother's age  Father's age  \
4             Yes                4.720703          32.0           NaN
283            No                5.015183           NaN           NaN
304            No                4.688240          45.0           NaN
513            No                5.058617          48.0          57.0
553            No                4.966875          51.0          47.0

      Institute Name                                Location of Institute  …  \
4    Carney Hospital  300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42…  …
283   Not applicable                                                 -  …
304              NaN  300 LONGWOOD AV\nFENWAY/KENMORE, MA 02115\n(42…  …
513   Not applicable                                                 -  …
553              NaN  818 HARRISON AV\nSOUTH END, MA 02118\n(42.3359…  …

    Birth defects White Blood cell count (thousand per microliter)  \
4        Multiple                                          4.098210
283           NaN                                          3.480522
304           NaN                                         11.531087
513      Singular                                          6.327895
553      Multiple                                          6.930982

    Blood test result  Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
4                 NaN         0.0         0.0         0.0         0.0        NaN
283          abnormal         NaN         0.0         NaN         0.0        0.0
304      inconclusive         0.0         NaN         0.0         0.0        0.0
513            normal         0.0         0.0         NaN         0.0        0.0
553          abnormal         1.0         0.0         0.0         0.0        0.0

                                Genetic Disorder Disorder Subclass
4    Multifactorial genetic inheritance disorders           Cancer
283  Multifactorial genetic inheritance disorders           Cancer
304  Multifactorial genetic inheritance disorders           Cancer
513  Multifactorial genetic inheritance disorders           Cancer
553  Multifactorial genetic inheritance disorders           Cancer

[5 rows x 41 columns]
```

[121]:
```
cm = cancer_and_mental_illness
cm_num = cm.select_dtypes(exclude='object')
cm_num.head()
```

[121]:
```
     Patient Age  Blood cell count (mcL)  Mother's age  Father's age  Test 1  \
4           11.0                4.720703          32.0           NaN     0.0
283         12.0                5.015183           NaN           NaN     NaN
304          4.0                4.688240          45.0           NaN     0.0
513          2.0                5.058617          48.0          57.0     0.0
553          7.0                4.966875          51.0          47.0     0.0

     Test 2  Test 3  Test 4  Test 5  No. of previous abortion  \
4       0.0     0.0     1.0     0.0                       4.0
283     0.0     0.0     1.0     0.0                       1.0
304     0.0     0.0     NaN     NaN                       3.0
513     0.0     0.0     1.0     0.0                       4.0
553     0.0     NaN     1.0     0.0                       0.0

     White Blood cell count (thousand per microliter)  Symptom 1  Symptom 2  \
4                                            4.098210        0.0        0.0
283                                          3.480522        NaN        0.0
304                                         11.531087        0.0        NaN
513                                          6.327895        0.0        0.0
553                                          6.930982        1.0        0.0

     Symptom 3  Symptom 4  Symptom 5
4          0.0        0.0        NaN
283        NaN        0.0        0.0
304        0.0        0.0        0.0
513        NaN        0.0        0.0
553        0.0        0.0        0.0
```

[122]:
```
cm_num['White Blood cell count (thousand per microliter)'].value_counts()
```

[122]:
```
White Blood cell count (thousand per microliter)
12.000000    6
3.000000     6
4.098210     1
6.470661     1
7.745894     1
            ..
9.694821     1
7.357842     1
6.452673     1
5.969282     1
7.100736     1
Name: count, Length: 74, dtype: int64
```

[123]:
```
df_disorders
```

[123]:
```
Disorder Subclass                    Genetic Disorder
Alzheimer's                          Multifactorial genetic inheritance
disorders     133
Cancer                               Multifactorial genetic inheritance
disorders      91
Cystic fibrosis                      Single-gene inheritance diseases
3145
Diabetes                             Multifactorial genetic inheritance
disorders    1653
Hemochromatosis                      Single-gene inheritance diseases
1228
Leber's hereditary optic neuropathy  Mitochondrial genetic inheritance disorders
587
Leigh syndrome                       Mitochondrial genetic inheritance disorders
4683
Mitochondrial myopathy               Mitochondrial genetic inheritance disorders
3971
Tay-Sachs                            Single-gene inheritance diseases
2556
Name: count, dtype: int64
```

[124]:
```
diabetes = df[df['Disorder Subclass'] == 'Diabetes']
diabetes.head()
```

[124]:

	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Mother's age	Father's age	Institute Name	Location of Institute	…	Birth defects
2	6.0	Yes	No	No	No	4.893297	41.0	22.0	NaN	-	…	Singular
9	4.0	No	Yes	Yes	Yes	4.752272	44.0	42.0	Shriners Burns Institute			
37	10.0	Yes	Yes	Yes	No	4.828440	51.0	NaN	Massachusetts Eye & Ear Infirmary			
58	5.0	Yes	Yes	Yes	Yes	4.771483	47.0	NaN	Not applicable			
77	5.0	No	No	No	Yes	4.851361	42.0	34.0	Massachusetts Eye & Ear Infirmary			

```
9   1200 Centre St\nRoslindale, MA 02131\n(42.2973…  …      Multiple
37  49 ROBINWOOD AV\nJAMAICA PLAIN, MA 02130\n(42…  …      Singular
58                                                  -  …      Multiple
77  59 TOWNSEND ST\nROXBURY, MA 02119\n(42.3185628…  …      Singular

   White Blood cell count (thousand per microliter)  Blood test result  \
2                                               NaN             normal
9                                          6.397702           abnormal
37                                         4.829049  slightly abnormal
58                                        10.682594  slightly abnormal
77                                         6.097961       inconclusive

    Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
2         0.0        1.0        1.0        1.0        1.0
9         0.0        0.0        1.0        1.0        1.0
37        1.0        1.0        1.0        NaN        1.0
58        NaN        1.0        NaN        1.0        1.0
77        1.0        1.0        1.0        1.0        1.0

                                Genetic Disorder Disorder Subclass
2   Multifactorial genetic inheritance disorders          Diabetes
9   Multifactorial genetic inheritance disorders          Diabetes
37  Multifactorial genetic inheritance disorders          Diabetes
58  Multifactorial genetic inheritance disorders          Diabetes
77  Multifactorial genetic inheritance disorders          Diabetes

[5 rows x 41 columns]
```

[131]:
```
# symptoms and diabetes

diabetes_symptoms = diabetes[['Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom␣
 ↪4',
       'Symptom 5']]
ds = diabetes_symptoms.value_counts()

diabetes_symptoms = diabetes_symptoms.dropna()
diabetes_symptoms
```

[131]:
```
       Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5
2            0.0        1.0        1.0        1.0        1.0
9            0.0        0.0        1.0        1.0        1.0
77           1.0        1.0        1.0        1.0        1.0
116          1.0        1.0        0.0        1.0        0.0
125          0.0        1.0        1.0        1.0        1.0
…            …          …          …          …          …
22040        1.0        1.0        1.0        1.0        0.0
22068        1.0        1.0        1.0        1.0        1.0
```

```
22070        0.0        1.0        1.0        1.0        1.0
22079        1.0        1.0        1.0        1.0        0.0
22082        1.0        0.0        1.0        1.0        1.0

[1191 rows x 5 columns]
```

[136]:
```
ds.describe()
```

[136]:
```
count     19.000000
mean      62.684211
std       71.978432
min        1.000000
25%       16.000000
50%       25.000000
75%      126.500000
max      248.000000
Name: count, dtype: float64
```

[138]:
```
diabetes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1817 entries, 2 to 22082
Data columns (total 41 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Patient Age                             1692 non-null   float64
 1   Genes in mother's side                  1817 non-null   object
 2   Inherited from father                   1794 non-null   object
 3   Maternal gene                           1595 non-null   object
 4   Paternal gene                           1817 non-null   object
 5   Blood cell count (mcL)                  1817 non-null   float64
 6   Mother's age                            1358 non-null   float64
 7   Father's age                            1362 non-null   float64
 8   Institute Name                          1427 non-null   object
 9   Location of Institute                   1817 non-null   object
 10  Status                                  1817 non-null   object
 11  Respiratory Rate (breaths/min)          1664 non-null   object
 12  Heart Rate (rates/min                   1663 non-null   object
 13  Test 1                                  1650 non-null   float64
 14  Test 2                                  1653 non-null   float64
 15  Test 3                                  1635 non-null   float64
 16  Test 4                                  1649 non-null   float64
 17  Test 5                                  1613 non-null   float64
 18  Parental consent                        1645 non-null   object
 19  Follow-up                               1649 non-null   object
 20  Gender                                  1661 non-null   object
 21  Birth asphyxia                          1663 non-null   object
```

```
 22  Autopsy shows birth defect (if applicable)        1476 non-null   object
 23  Place of birth                                    1662 non-null   object
 24  Folic acid details (peri-conceptional)            1640 non-null   object
 25  H/O serious maternal illness                      1660 non-null   object
 26  H/O radiation exposure (x-ray)                    1648 non-null   object
 27  H/O substance abuse                               1648 non-null   object
 28  Assisted conception IVF/ART                       1650 non-null   object
 29  History of anomalies in previous pregnancies      1639 non-null   object
 30  No. of previous abortion                          1661 non-null   float64
 31  Birth defects                                     1654 non-null   object
 32  White Blood cell count (thousand per microliter)  1651 non-null   float64
 33  Blood test result                                 1663 non-null   object
 34  Symptom 1                                         1638 non-null   float64
 35  Symptom 2                                         1648 non-null   float64
 36  Symptom 3                                         1668 non-null   float64
 37  Symptom 4                                         1661 non-null   float64
 38  Symptom 5                                         1633 non-null   float64
 39  Genetic Disorder                                  1653 non-null   object
 40  Disorder Subclass                                 1817 non-null   object
dtypes: float64(16), object(25)
memory usage: 596.2+ KB
```

[140]:
```
df_disorders
```

[140]:
```
Disorder Subclass                   Genetic Disorder
Alzheimer's                         Multifactorial genetic inheritance
disorders     133
Cancer                              Multifactorial genetic inheritance
disorders      91
Cystic fibrosis                     Single-gene inheritance diseases
3145
Diabetes                            Multifactorial genetic inheritance
disorders    1653
Hemochromatosis                     Single-gene inheritance diseases
1228
Leber's hereditary optic neuropathy  Mitochondrial genetic inheritance disorders
587
Leigh syndrome                      Mitochondrial genetic inheritance disorders
4683
Mitochondrial myopathy              Mitochondrial genetic inheritance disorders
3971
Tay-Sachs                           Single-gene inheritance diseases
2556
Name: count, dtype: int64
```

[141]:
```
# hemochromatosis
```

```
hemo = df[df['Disorder Subclass'] == 'Hemochromatosis']
hemo.head()
```

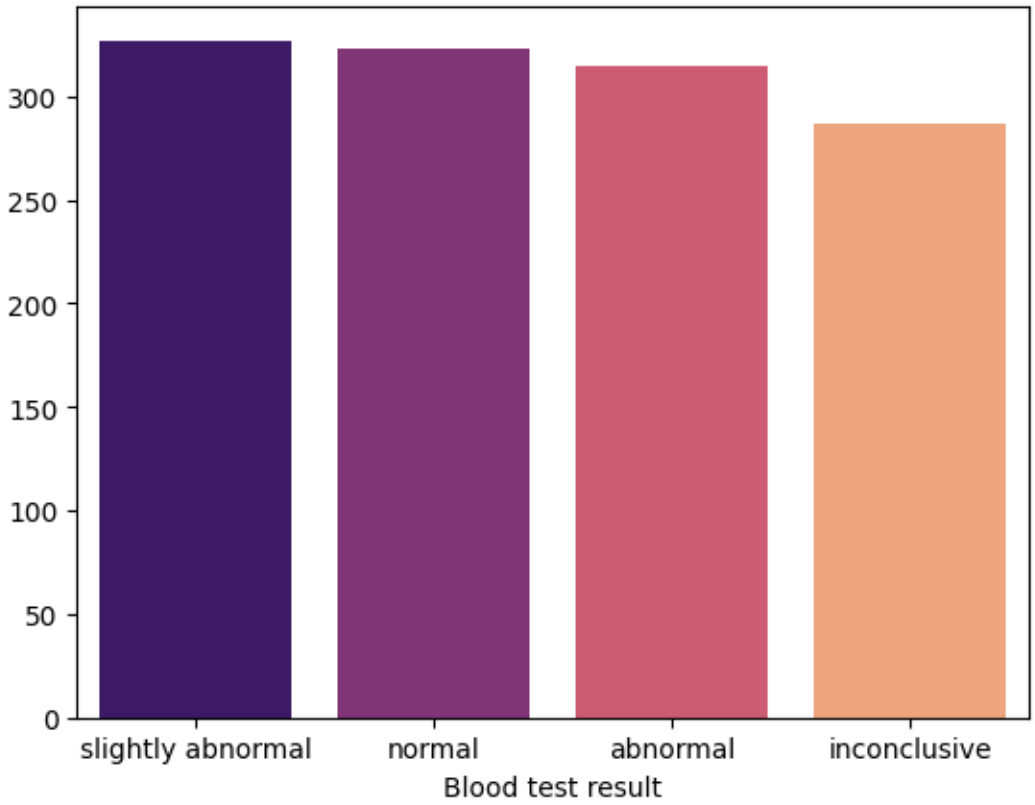
[141]:
```
    Patient Age Genes in mother's side Inherited from father Maternal gene  \
10          6.0                    Yes                    No           NaN
19          6.0                     No                   Yes           Yes
20          2.0                     No                    No           Yes
44          9.0                    Yes                    No           Yes
50          9.0                     No                   Yes            No

   Paternal gene  Blood cell count (mcL)  Mother's age  Father's age  \
10            No                4.750824           NaN           NaN
19           Yes                4.876896          36.0          48.0
20            No                4.808872           NaN          30.0
44            No                4.970435          50.0          51.0
50           Yes                5.028235          30.0          50.0

                Institute Name  \
10              Not applicable
19                 VA Hospital
20              Not applicable
44              Not applicable
50  New England Baptist Hospital

                             Location of Institute  … Birth defects  \
10                                               -  …      Singular
19  249 RIVER ST\nMATTAPAN, MA 02126\n(42.27137912…  …      Singular
20                                               -  …           NaN
44                                               -  …      Multiple
50  125 NASHUA ST\nCENTRAL, MA 02114\n(42.36764789…  …      Singular

   White Blood cell count (thousand per microliter)  Blood test result  \
10                                          5.957321           abnormal
19                                          7.370477             normal
20                                          9.566103  slightly abnormal
44                                         11.648665           abnormal
50                                          7.237478           abnormal

    Symptom 1  Symptom 2  Symptom 3  Symptom 4  Symptom 5  \
10        1.0        NaN        0.0        0.0        NaN
19        1.0        0.0        0.0        0.0        0.0
20        1.0        0.0        0.0        1.0        0.0
44        0.0        1.0        NaN        0.0        0.0
50        0.0        0.0        0.0        0.0        1.0

                  Genetic Disorder Disorder Subclass
10  Single-gene inheritance diseases   Hemochromatosis
```

```
19  Single-gene inheritance diseases   Hemochromatosis
20  Single-gene inheritance diseases   Hemochromatosis
44  Single-gene inheritance diseases   Hemochromatosis
50  Single-gene inheritance diseases   Hemochromatosis

[5 rows x 41 columns]
```

[144]:
```
hemo_blood_test = hemo['Blood test result'].value_counts()
```

[146]:
```
sns.barplot(x=hemo_blood_test.index, y=hemo_blood_test.values, palette='magma')
```

[146]:
```
<Axes: xlabel='Blood test result'>
```



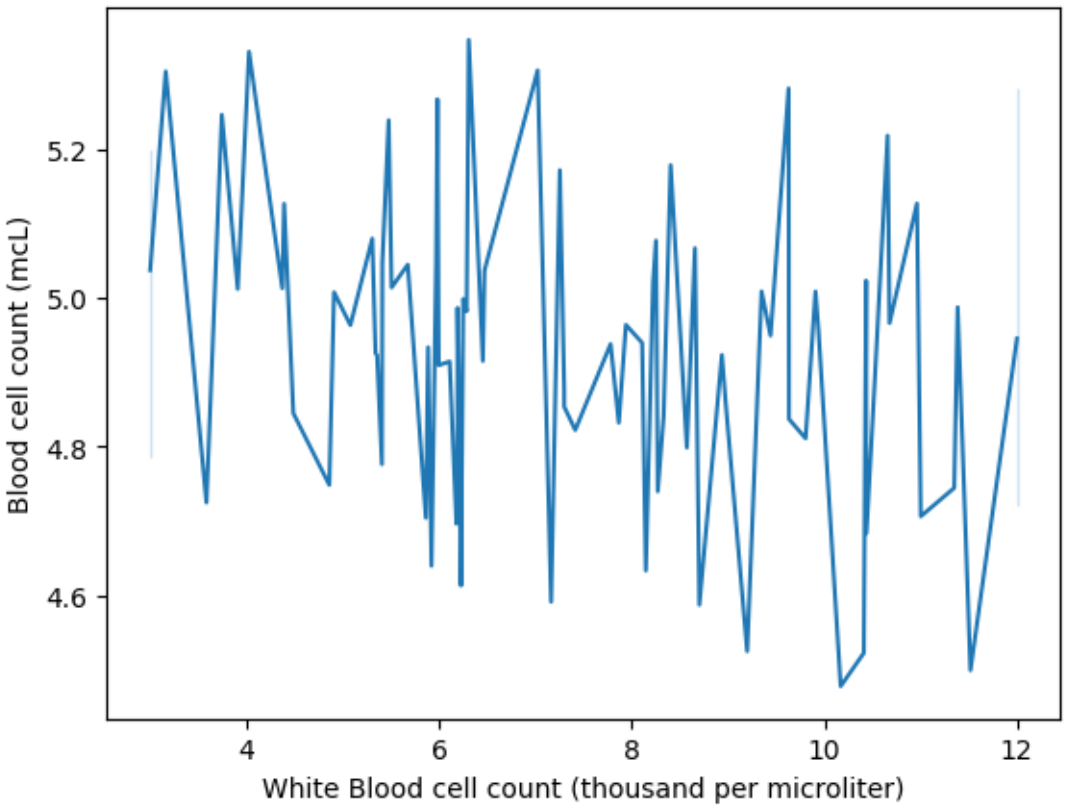
[152]:
```
white_blood_sample = hemo['White Blood cell count (thousand per microliter)'].
↪sample(100)
white_blood_sample

df_wbs = pd.DataFrame(white_blood_sample)
df_wbs
```

[152]:
```
      White Blood cell count (thousand per microliter)
6398                                                NaN
21177                                          9.231183
15524                                          3.000000
9207                                           3.000000
12649                                               NaN
…                                                   …
5829                                           5.534540
14599                                          7.364213
8013                                                NaN
13602                                         12.000000
12320                                          3.000000

[100 rows x 1 columns]
```

[158]:
```
df_samp = df.sample(100)
```

[159]:
```
sns.lineplot(x='White Blood cell count (thousand per microliter)',y='Blood cell␣
↪count (mcL)', data=df_samp, palette='crest')
```

[159]:
```
<Axes: xlabel='White Blood cell count (thousand per microliter)', ylabel='Blood
cell count (mcL)'>
```



[ ]: