eda-report

July 3, 2024

[119]:

```
import pandas as pd
df = pd.read_csv('lung_cancer_data.csv')
df.head(10)
```

[119]:

```
    Patient_ID  Age  Gender Smoking_History  Tumor_Size_mm Tumor_Location  \
0  Patient0000   68    Male  Current Smoker      81.678677     Lower Lobe
1  Patient0001   58    Male    Never Smoked      78.448272     Lower Lobe
2  Patient0002   44    Male   Former Smoker      67.714305     Lower Lobe
3  Patient0003   72    Male  Current Smoker      70.806008     Lower Lobe
4  Patient0004   37  Female    Never Smoked      87.272433     Lower Lobe
5  Patient0005   50    Male    Never Smoked      72.148656     Lower Lobe
6  Patient0006   68  Female  Current Smoker      19.122175    Middle Lobe
7  Patient0007   48    Male  Current Smoker      68.095057     Lower Lobe
8  Patient0008   52  Female   Former Smoker      25.299440     Lower Lobe
9  Patient0009   40    Male  Current Smoker      11.282767     Lower Lobe

       Stage          Treatment  Survival_Months          Ethnicity  …  \
0  Stage III            Surgery               44           Hispanic  …
1    Stage I  Radiation Therapy              101          Caucasian  …
2    Stage I       Chemotherapy               69   African American  …
3  Stage III       Chemotherapy               95   African American  …
4   Stage IV  Radiation Therapy              105              Asian  …
5    Stage I            Surgery               49           Hispanic  …
6    Stage I  Radiation Therapy               63   African American  …
7   Stage IV       Chemotherapy              101   African American  …
8    Stage I   Targeted Therapy               35          Caucasian  …
9    Stage I            Surgery               19              Other  …

  Alanine_Aminotransferase_Level Aspartate_Aminotransferase_Level  \
0                      27.985571                         46.801214
1                      30.120956                         39.711531
2                       5.882418                         32.640602
3                      38.908154                         44.319393
4                      26.344877                         15.746906
5                      34.813869                         29.769655
6                      31.016446                         39.878953
7                      12.208267                         23.908107
```

```
8                      36.888358                        35.822953
9                      33.836074                        44.230240

  Creatinine_Level   LDH_Level Calcium_Level Phosphorus_Level Glucose_Level  \
0         1.245849  239.240255     10.366307         3.547734    113.919243
1         1.463231  233.515237     10.081731         2.945020    101.321578
2         0.630109  169.037460      8.660892         4.637399     78.214177
3         0.594342  213.967590      8.832669         3.617098    127.895361
4         1.478239  118.187543      9.247609         4.773255    148.801185
5         0.825544  218.204614      8.711924         2.661053    142.782619
6         0.799593  181.550728      8.089885         4.591886     75.377094
7         1.436453  119.057097      9.367766         4.909359     99.511881
8         1.089169  197.791757     10.188013         3.326973    145.657154
9         1.078794  227.048430      8.248718         3.173471    109.755478

  Potassium_Level Sodium_Level  Smoking_Pack_Years
0        4.968163   139.822861           17.006956
1        3.896795   135.449361           93.270893
2        4.369050   143.377155           70.348376
3        4.348474   138.586005           19.828128
4        3.671976   141.230724           81.047456
5        4.606625   135.497944           18.058525
6        4.800980   138.373413           86.482339
7        4.061255   136.347159           68.239920
8        4.767092   141.113503           96.808889
9        4.075269   139.174855           68.595875

[10 rows x 38 columns]
```

[120]:

```
df = df.drop(columns=['Patient_ID','Performance_Status'])
```

[121]:

```
df.isnull().sum()
```

[121]:

```
Age                                 0
Gender                              0
Smoking_History                     0
Tumor_Size_mm                       0
Tumor_Location                      0
Stage                               0
Treatment                           0
Survival_Months                     0
Ethnicity                           0
Insurance_Type                      0
Family_History                      0
Comorbidity_Diabetes                0
Comorbidity_Hypertension            0
Comorbidity_Heart_Disease           0
```

```
Comorbidity_Chronic_Lung_Disease    0
Comorbidity_Kidney_Disease          0
Comorbidity_Autoimmune_Disease      0
Comorbidity_Other                   0
Blood_Pressure_Systolic             0
Blood_Pressure_Diastolic            0
Blood_Pressure_Pulse                0
Hemoglobin_Level                    0
White_Blood_Cell_Count              0
Platelet_Count                      0
Albumin_Level                       0
Alkaline_Phosphatase_Level          0
Alanine_Aminotransferase_Level      0
Aspartate_Aminotransferase_Level    0
Creatinine_Level                    0
LDH_Level                           0
Calcium_Level                       0
Phosphorus_Level                    0
Glucose_Level                       0
Potassium_Level                     0
Sodium_Level                        0
Smoking_Pack_Years                  0
dtype: int64
```

[122]:

```
df.describe()
```

[122]:

```
                Age  Tumor_Size_mm  Survival_Months  Blood_Pressure_Systolic  \
count  23658.000000   23658.000000     23658.000000             23658.000000
mean      54.439344      55.383736        59.863809               134.462381
std       14.396386      26.004354        34.246042                26.020492
min       30.000000      10.004279         1.000000                90.000000
25%       42.000000      32.972797        30.000000               112.000000
50%       54.000000      55.296297        60.000000               134.000000
75%       67.000000      78.190014        89.000000               157.000000
max       79.000000      99.990554       119.000000               179.000000

       Blood_Pressure_Diastolic  Blood_Pressure_Pulse  Hemoglobin_Level  \
count              23658.000000          23658.000000      23658.000000
mean                  84.475780             79.585299         14.000137
std                   14.409826             11.546690          2.301411
min                   60.000000             60.000000         10.000070
25%                   72.000000             70.000000         11.990625
50%                   85.000000             80.000000         13.983383
75%                   97.000000             90.000000         15.999260
max                  109.000000             99.000000         17.999957

       White_Blood_Cell_Count  Platelet_Count  Albumin_Level  …  \
count            23658.000000    23658.000000   23658.000000  …
mean                 6.735637      299.867482       3.998981  …
std                  1.879292       86.897568       0.576931  …
min                  3.501213      150.017892       3.000080  …
25%                  5.108723      224.884576       3.504579  …
50%                  6.729774      299.933443       3.999931  …
75%                  8.353701      375.437029       4.499102  …
max                  9.999535      449.974734       4.999968  …

       Alanine_Aminotransferase_Level  Aspartate_Aminotransferase_Level  \
count                    23658.000000                      23658.000000
mean                        22.504677                         30.133226
std                         10.047864                         11.560915
min                          5.001090                         10.000860
25%                         13.816180                         20.065339
50%                         22.547943                         30.271772
75%                         31.092935                         40.107488
max                         39.999543                         49.998571

       Creatinine_Level     LDH_Level  Calcium_Level  Phosphorus_Level  \
count      23658.000000  23658.000000   23658.000000      23658.000000
mean           0.999459    174.734575       9.261114          3.742771
std            0.287517     43.230997       0.719875          0.721708
min            0.500001    100.002721       8.000018          2.500069
25%            0.748845    137.444977       8.640877          3.120107
50%            1.001183    174.390634       9.259304          3.730837
75%            1.249173    212.228273       9.883248          4.364422
max            1.499998    249.996391      10.499913          4.999974

       Glucose_Level  Potassium_Level  Sodium_Level  Smoking_Pack_Years
count   23658.000000     23658.000000  23658.000000        23658.000000
mean      109.895553         4.245646    140.028215           49.913594
std        23.109136         0.431968      2.894568           28.870940
min        70.000420         3.500034    135.000934            0.016800
25%        89.828616         3.871842    137.540078           25.026793
50%       109.949488         4.242236    140.002209           49.926220
75%       130.061977         4.618318    142.541883           74.924580
max       149.997056         4.999954    144.999869           99.999493

[8 rows x 21 columns]
```

[123]:

```
#converting categorical variables into numerical variables

from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()
```

```python
df['Family_History'] = encoder.fit_transform(df['Family_History'])

print(df['Family_History'].unique())
```

```
[0 1]
```

[124]:

```python
df['Gender'] = encoder.fit_transform(df['Gender'])
```

[125]:

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23658 entries, 0 to 23657
Data columns (total 36 columns):
 #   Column                             Non-Null Count  Dtype  
---  ------                             --------------  -----  
 0   Age                                23658 non-null  int64  
 1   Gender                             23658 non-null  int32  
 2   Smoking_History                    23658 non-null  object 
 3   Tumor_Size_mm                      23658 non-null  float64
 4   Tumor_Location                     23658 non-null  object 
 5   Stage                              23658 non-null  object 
 6   Treatment                          23658 non-null  object 
 7   Survival_Months                    23658 non-null  int64  
 8   Ethnicity                          23658 non-null  object 
 9   Insurance_Type                     23658 non-null  object 
 10  Family_History                     23658 non-null  int32  
 11  Comorbidity_Diabetes               23658 non-null  object 
 12  Comorbidity_Hypertension           23658 non-null  object 
 13  Comorbidity_Heart_Disease          23658 non-null  object 
 14  Comorbidity_Chronic_Lung_Disease   23658 non-null  object 
 15  Comorbidity_Kidney_Disease         23658 non-null  object 
 16  Comorbidity_Autoimmune_Disease     23658 non-null  object 
 17  Comorbidity_Other                  23658 non-null  object 
 18  Blood_Pressure_Systolic            23658 non-null  int64  
 19  Blood_Pressure_Diastolic           23658 non-null  int64  
 20  Blood_Pressure_Pulse               23658 non-null  int64  
 21  Hemoglobin_Level                   23658 non-null  float64
 22  White_Blood_Cell_Count             23658 non-null  float64
 23  Platelet_Count                     23658 non-null  float64
 24  Albumin_Level                      23658 non-null  float64
 25  Alkaline_Phosphatase_Level         23658 non-null  float64
 26  Alanine_Aminotransferase_Level     23658 non-null  float64
 27  Aspartate_Aminotransferase_Level   23658 non-null  float64
 28  Creatinine_Level                   23658 non-null  float64
 29  LDH_Level                          23658 non-null  float64
 30  Calcium_Level                      23658 non-null  float64
 31  Phosphorus_Level                   23658 non-null  float64
 32  Glucose_Level                      23658 non-null  float64
 33  Potassium_Level                    23658 non-null  float64
 34  Sodium_Level                       23658 non-null  float64
 35  Smoking_Pack_Years                 23658 non-null  float64
dtypes: float64(16), int32(2), int64(5), object(13)
memory usage: 6.3+ MB
```

[126]:

```
df['Smoking_History'] = encoder.fit_transform(df['Smoking_History'])
df['Smoking_History'].unique()
```

[126]:

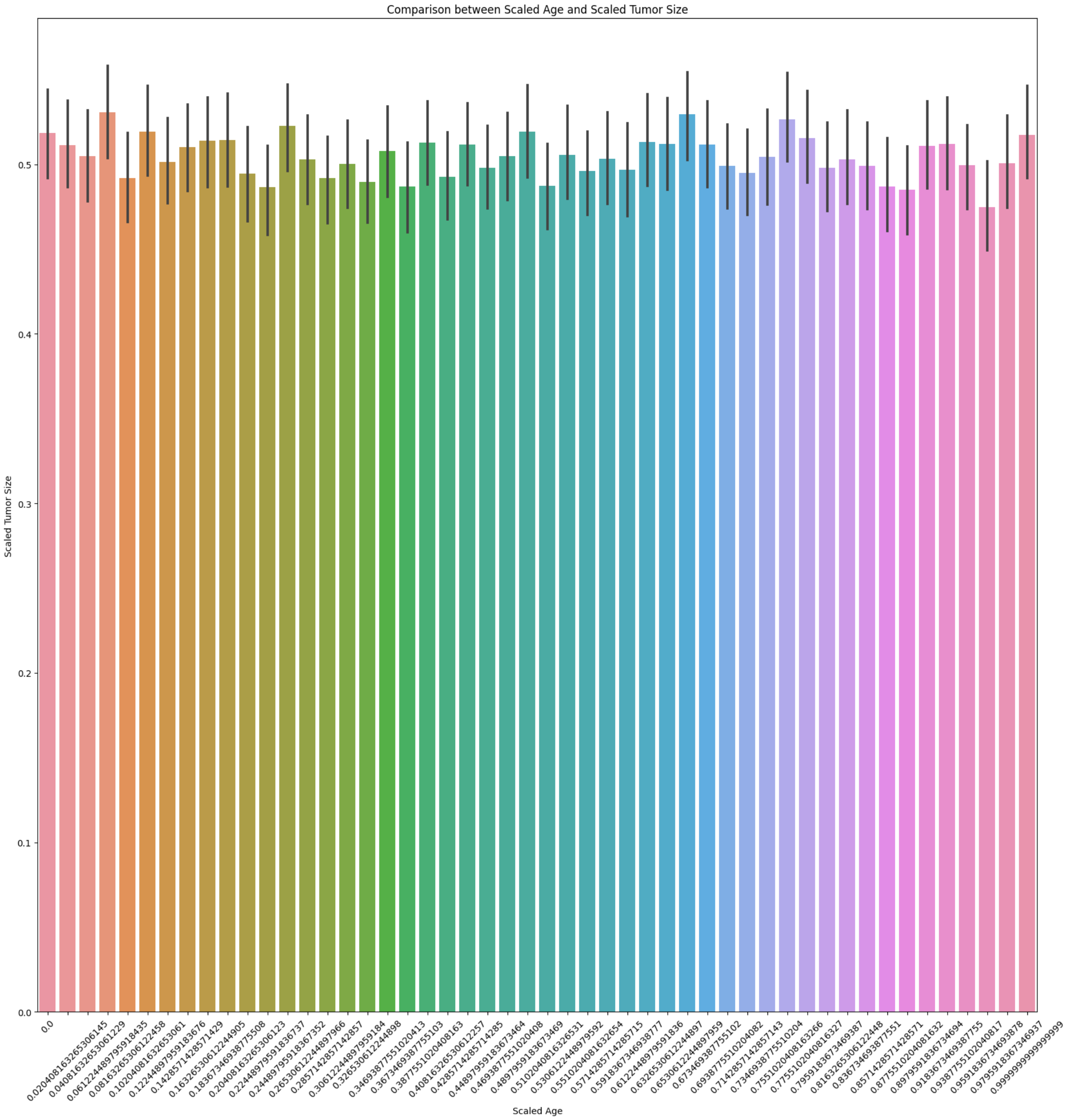
```
array([0, 2, 1])
```

[127]:

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()

scaled_size= scaler.fit_transform(df['Tumor_Size_mm'].values.reshape(-1, 1))
scaled_age = scaler.fit_transform(df['Age'].values.reshape(-1,1))
```

[128]:

```
# Comparing between Age and Tumor Size
import seaborn as sns
import matplotlib.pyplot as plt
scaled_df = pd.DataFrame({'Age': scaled_age.flatten(), 'Tumor_Size_mm':␣
 ↪scaled_size.flatten()})
plt.figure(figsize=(20, 20))
sns.barplot(x='Age', y='Tumor_Size_mm', data=scaled_df)
plt.xlabel('Scaled Age')
plt.ylabel('Scaled Tumor Size')
plt.title('Comparison between Scaled Age and Scaled Tumor Size')
plt.xticks(rotation=45)
plt.show()
```



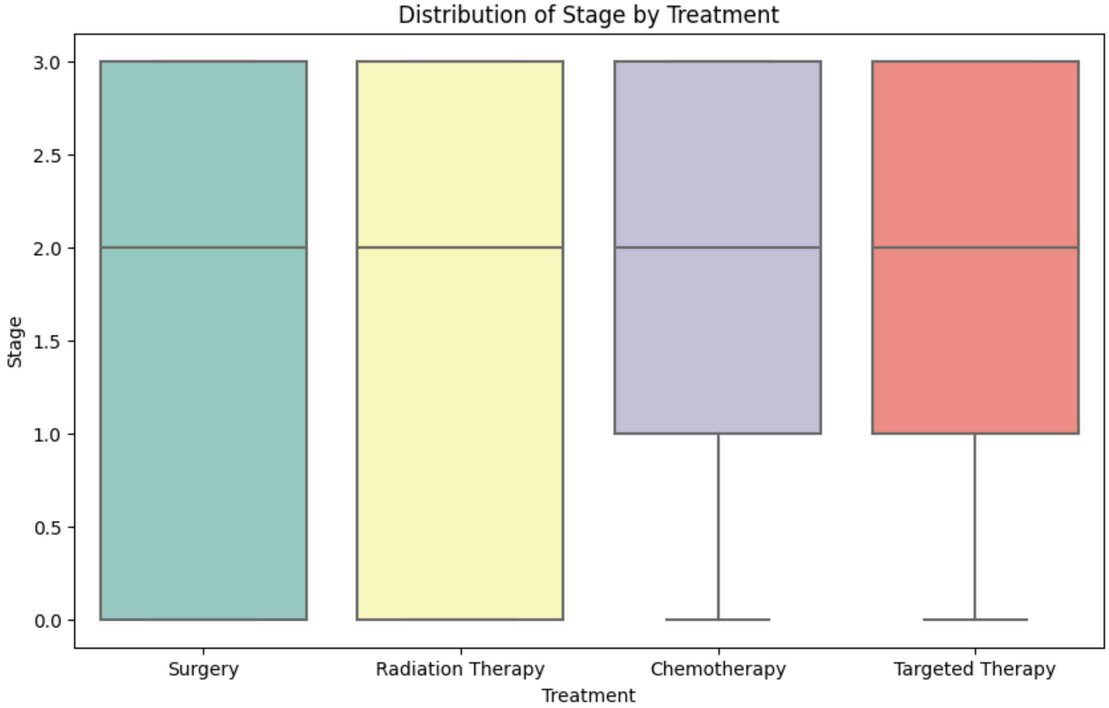
[129]:
```
#Converting stage to numerical
df['Stage'] = encoder.fit_transform(df['Stage'])
df['Stage'].unique()
```

[129]:
```
array([2, 0, 3, 1])
```

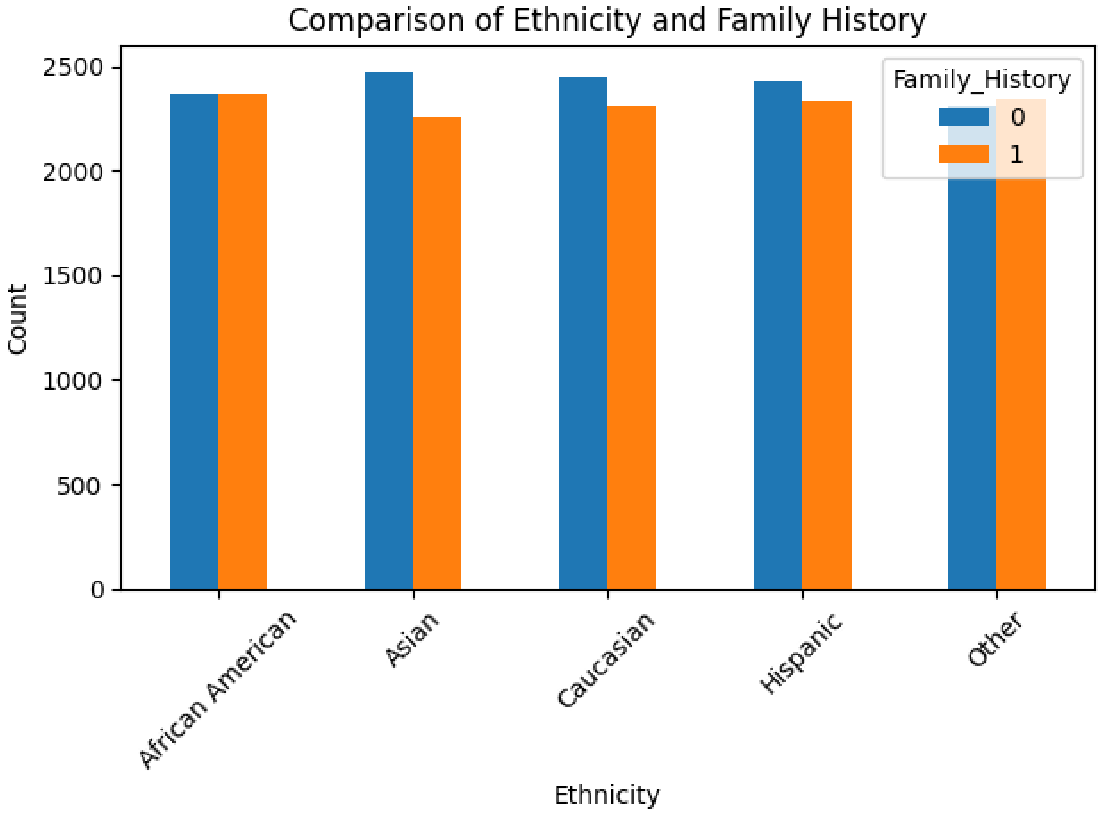
[130]:
```
#Stage vs Treatment

df['Treatment'].unique()
```

[130]:
```
array(['Surgery', 'Radiation Therapy', 'Chemotherapy', 'Targeted Therapy'],
      dtype=object)
```

[131]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Treatment', y='Stage', data=df, palette='Set3')
plt.title('Distribution of Stage by Treatment')
plt.xlabel('Treatment')
plt.ylabel('Stage')
plt.show()
```

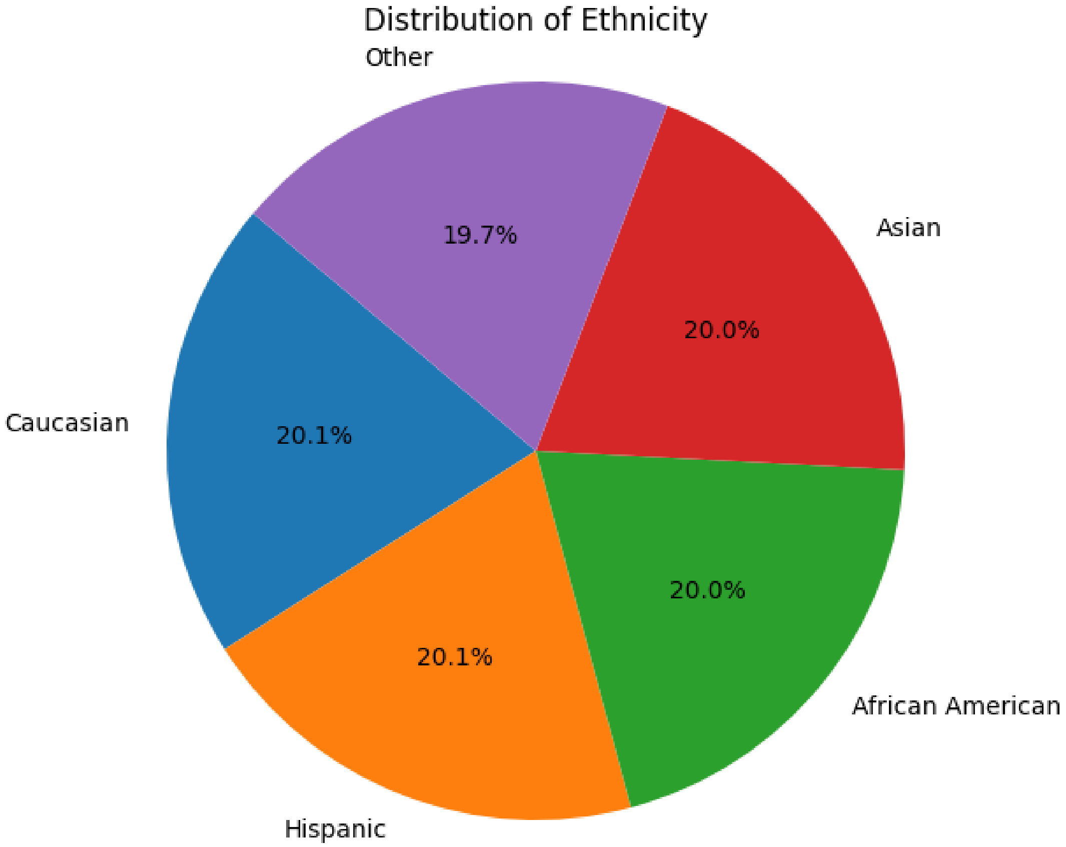


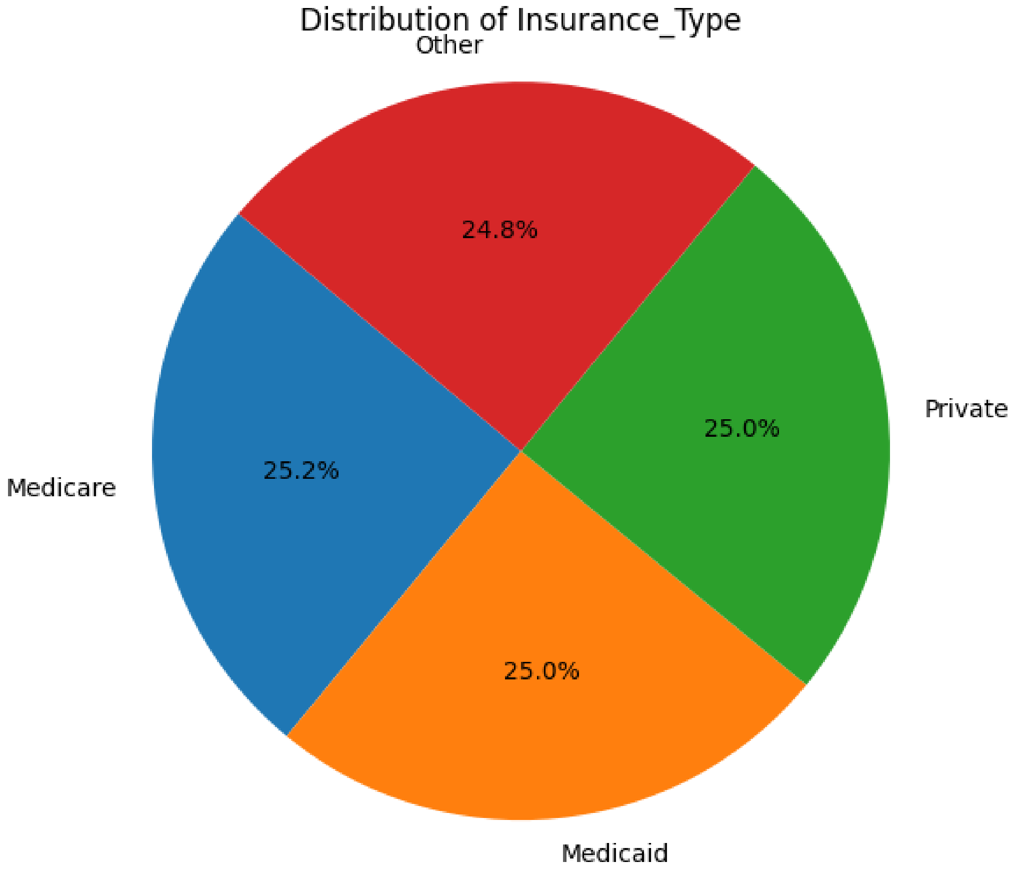
[132]:
```
# Group by Ethnicity and Family_History to calculate frequencies
grouped = df.groupby(['Ethnicity', 'Family_History']).size().
 ↪reset_index(name='count')

# Pivot the data for plotting
pivot_data = grouped.pivot(index='Ethnicity', columns='Family_History',␣
 ↪values='count').fillna(0)

# Plotting
pivot_data.plot(kind='bar', stacked=False)
plt.title('Comparison of Ethnicity and Family History')
plt.xlabel('Ethnicity')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

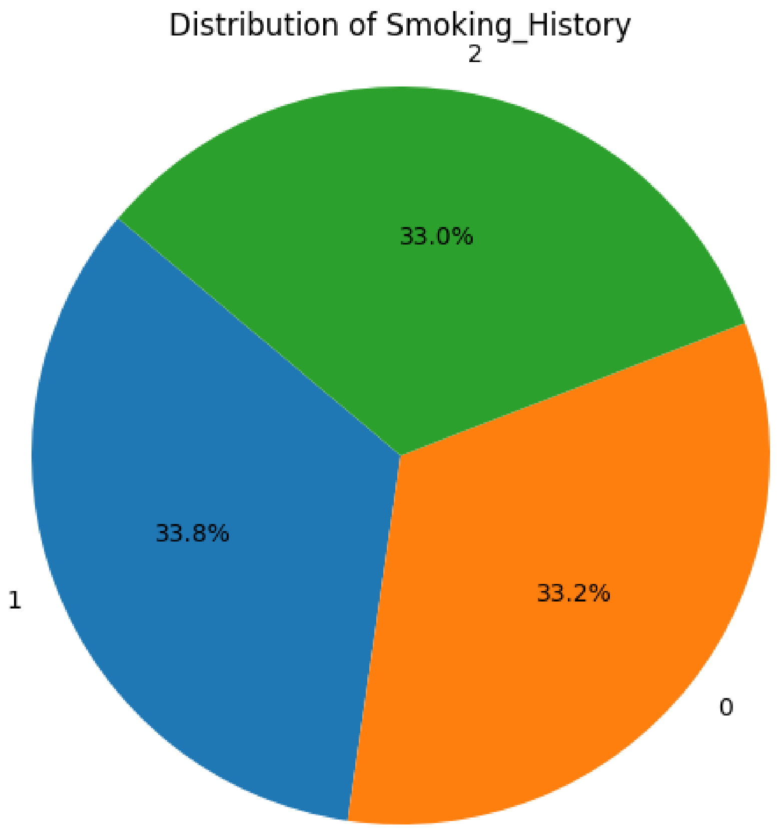


[133]:

```
def create_pie_chart(column_name):
    counts = df[column_name].value_counts()
    plt.figure(figsize=(8, 6))
    plt.pie(counts, labels=counts.index, autopct='%1.1f%%', startangle=140)
    plt.title(f'Distribution of {column_name}')
    plt.axis('equal')
    plt.show()

# Create pie charts for each categorical variable
create_pie_chart('Ethnicity')
create_pie_chart('Insurance_Type')
create_pie_chart('Smoking_History')
```

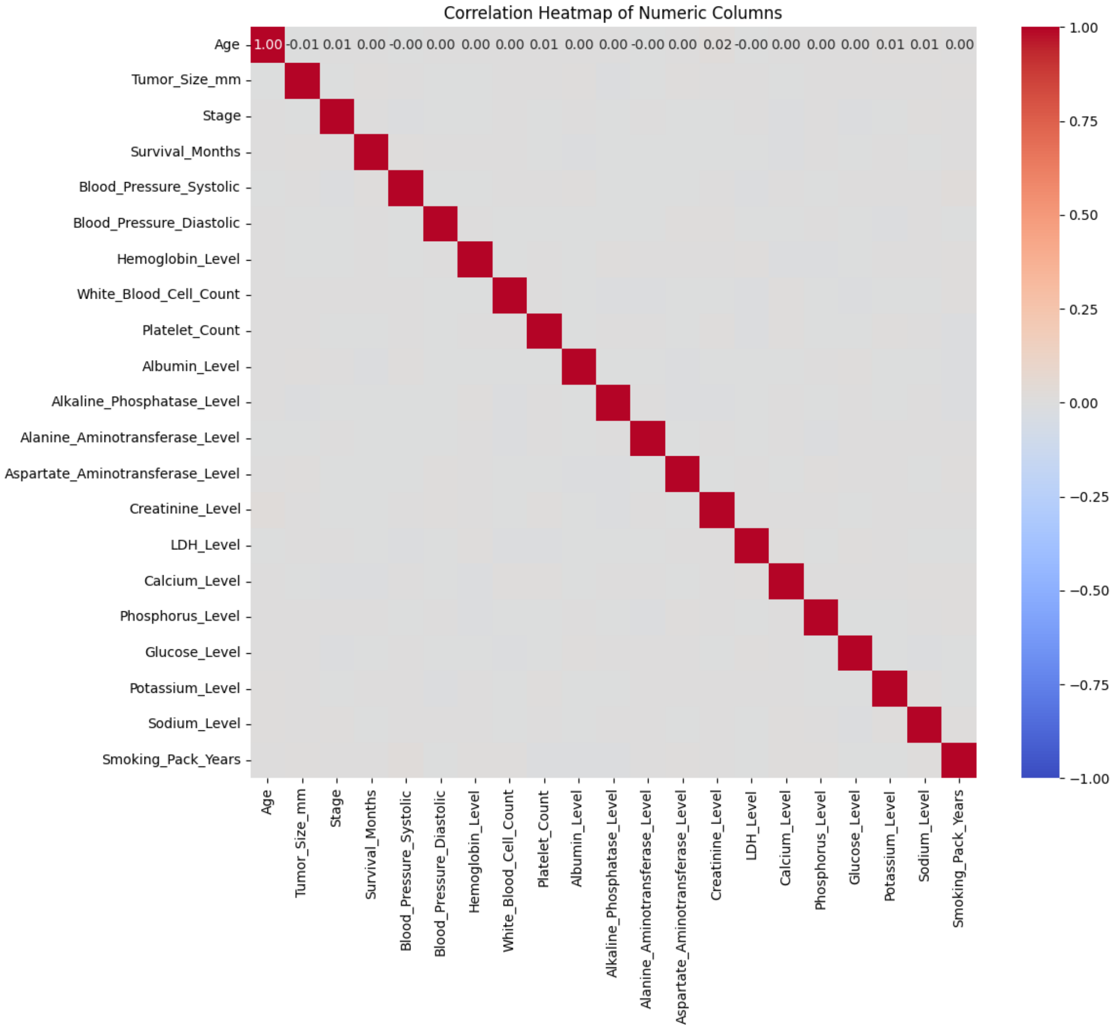

Distribution of Ethnicity
Other
19.7%
Asian
20.0%
Caucasian
20.1%
20.0%
20.1%
African American
Hispanic


Distribution of Insurance_Type
Other
24.8%
Private
25.0%
Medicare
25.2%
25.0%
Medicaid



[135]:

```
numeric_columns = [
    'Age', 'Tumor_Size_mm', 'Stage', 'Survival_Months',␣
 ↪'Blood_Pressure_Systolic',
    'Blood_Pressure_Diastolic', 'Hemoglobin_Level', 'White_Blood_Cell_Count',
    'Platelet_Count', 'Albumin_Level', 'Alkaline_Phosphatase_Level',
    'Alanine_Aminotransferase_Level', 'Aspartate_Aminotransferase_Level',
    'Creatinine_Level', 'LDH_Level', 'Calcium_Level', 'Phosphorus_Level',
    'Glucose_Level', 'Potassium_Level', 'Sodium_Level', 'Smoking_Pack_Years'
]

corr_matrix = df[numeric_columns].corr()

# Plotting the heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', vmin=-1,␣
 ↪vmax=1)
plt.title('Correlation Heatmap of Numeric Columns')
plt.show()
```


Correlation Heatmap of Numeric Columns
Age
Tumor_Size_mm
Stage
Survival_Months
Blood_Pressure_Systolic
Blood_Pressure_Diastolic
Hemoglobin_Level
White_Blood_Cell_Count
Platelet_Count
Albumin_Level
Alkaline_Phosphatase_Level
Alanine_Aminotransferase_Level
Aspartate_Aminotransferase_Level
Creatinine_Level
LDH_Level
Calcium_Level
Phosphorus_Level
Glucose_Level
Potassium_Level
Sodium_Level
Smoking_Pack_Years
1.00
-0.01
0.01
0.00
-0.00
0.00
0.00
0.00
0.01
0.00
0.00
-0.00
0.00
0.02
-0.00
0.00
0.00
0.00
0.01
0.01
0.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00