assignment-2

February 3, 2024

[40]:
```
import pandas as pd
```

[41]:
```
# 1. Load the data into a DataFrame
df = pd.read_csv("NCI_SEER_CRC.csv")
```

[42]:
```
# 2. Rename the columns
df.rename(columns={
    'Year of diagnosis': 'YDD',
    'Race recode (W, B, AI, API)': 'Race',
    'Origin recode NHIA (Hispanic, Non-Hisp)': 'Ethnicity',
    'Age recode with single ages and 100+': 'Age',
    'Grade (thru 2017)': 'Grade',
    'Marital status at diagnosis': 'Marital Status'
}, inplace=True)
```

[43]:
```
# 3. Drop the specified columns
sel_cols = [
    'Age recode with <1 year olds',
    'Primary Site - labeled',
    'Total number of in situ/malignant tumors for patient',
    'COD to site recode',
    'SEER registry (with CA and GA as whole states)',
    'Behavior recode for analysis'
]
df.drop(columns=sel_cols, inplace=True)
```

[44]:
```
# 4. Remove 'years' from the Age column
df['Age'] = df['Age'].str.replace(' years', '', regex=False)
```

[45]:
```
# 5. Enumerate Ethnicity
df['Ethnicity'] = df['Ethnicity'].map({'Spanish-Hispanic-Latino': 1,␣
↪'Non-Spanish-Hispanic-Latino': 0})
```

[46]:
```
# 6. Enumerate Grade
grds = {
    'Well differentiated; Grade I': 1,
    'Moderately differentiated; Grade II': 2,
```

```
    'Poorly differentiated; Grade III': 3,
    'Undifferentiated; anaplastic; Grade IV': 4,
    'Unknown': 99
}
df['Grade'] = df['Grade'].map(grds)
```

[47]:
```
# 7. Enumerate Sex
df['Sex'] = df['Sex'].map({'Female': 1, 'Male': 0})
```

[48]:
```
# 8. Drop records with null values and fill empty cells with 0
df.dropna(inplace=True)
df.fillna(0, inplace=True)

print(df.head())
```

```
    YDD   Race  Ethnicity  Sex Age  Grade Combined Summary Stage (2004+)  \
0  2000  White          0    0  82    2.0                         Blank(s)
1  2000  White          0    0  76    2.0                         Blank(s)
2  2000  White          0    1  65    2.0                         Blank(s)
3  2000  White          0    0  86    3.0                         Blank(s)
4  2000  White          0    1  82    2.0                         Blank(s)

  Survival months                  Marital Status  Appalachia
0              17  Married (including common law)           1
1              67                         Unknown           1
2              41                         Unknown           1
3              15                         Unknown           1
4               0                         Unknown           1
```