eso_can_det

December 19, 2024

[ ]:

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

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df = pd.read_csv("Esophageal_Dataset.csv")
```

[3]:
```
df.head()
```

[3]:
```
   Unnamed: 0 patient_barcode tissue_source_site patient_id  \
0           0    TCGA-2H-A9GF                 2H       A9GF
1           1    TCGA-2H-A9GG                 2H       A9GG
2           2    TCGA-2H-A9GH                 2H       A9GH
3           3    TCGA-2H-A9GI                 2H       A9GI
4           4    TCGA-2H-A9GJ                 2H       A9GJ

                     bcr_patient_uuid informed_consent_verified  \
0  0500F1A6-A528-43F3-B035-12D3B7C99C0F                      YES
1  70084008-697D-442D-8F74-C12F8F598570                      YES
2  606DC5B8-7625-42A6-A936-504EF25623A4                      YES
3  CEAF98F8-517E-457A-BF29-ACFE22893D49                      YES
4  EE47CD59-C8D8-4B1E-96DB-91C679E4106F                      YES

  icd_o_3_site icd_o_3_histology icd_10  \
0        C15.5            8140/3  C15.5
1        C15.5            8140/3  C15.5
2        C15.5            8140/3  C15.5
3        C15.5            8140/3  C15.5
4        C15.5            8140/3  C15.5

  tissue_prospective_collection_indicator  …  \
```

```
0                                      NO  …
1                                      NO  …
2                                      NO  …
3                                      NO  …
4                                      NO  …

  primary_pathology_lymph_node_examined_count  \
0                                         8.0
1                                        19.0
2                                        30.0
3                                         8.0
4                                        19.0

   primary_pathology_number_of_lymphnodes_positive_by_he  \
0                                                  7.0
1                                                  4.0
2                                                  1.0
3                                                  4.0
4                                                  0.0

  primary_pathology_number_of_lymphnodes_positive_by_ihc  \
0                                                  0.0
1                                                  0.0
2                                                  0.0
3                                                  0.0
4                                                  0.0

  primary_pathology_planned_surgery_status  \
0                                      NaN
1                                      NaN
2                                      NaN
3                                      NaN
4                                      NaN

   primary_pathology_treatment_prior_to_surgery  \
0                                           NaN
1                                           NaN
2                                           NaN
3                                           NaN
4                                           NaN

   primary_pathology_residual_tumor  \
0                                R1
1                                R1
2                                R0
3                                R0
4                                R0
```

```
  primary_pathology_karnofsky_performance_score  \
0                                           NaN
1                                           NaN
2                                           NaN
3                                           NaN
4                                           NaN

  primary_pathology_eastern_cancer_oncology_group  \
0                                             NaN
1                                             NaN
2                                             NaN
3                                             NaN
4                                             NaN

  primary_pathology_radiation_therapy primary_pathology_postoperative_rx_tx
0                                  NO                                    NO
1                                  NO                                    NO
2                                  NO                                    NO
3                                  NO                                    NO
4                                  NO                                    NO

[5 rows x 85 columns]
```

[5]:
```
df.shape
```

[5]:
```
(3985, 85)
```

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

[7]:
```
Index(['Unnamed: 0', 'patient_barcode', 'tissue_source_site', 'patient_id',
       'bcr_patient_uuid', 'informed_consent_verified', 'icd_o_3_site',
       'icd_o_3_histology', 'icd_10',
       'tissue_prospective_collection_indicator',
       'tissue_retrospective_collection_indicator', 'days_to_birth',
       'country_of_birth', 'gender', 'height', 'weight',
       'country_of_procurement', 'state_province_of_procurement',
       'city_of_procurement', 'race_list', 'ethnicity', 'other_dx',
       'history_of_neoadjuvant_treatment', 'person_neoplasm_cancer_status',
       'vital_status', 'days_to_last_followup', 'days_to_death',
       'tobacco_smoking_history', 'age_began_smoking_in_years',
       'stopped_smoking_year', 'number_pack_years_smoked',
       'alcohol_history_documented', 'frequency_of_alcohol_consumption',
       'amount_of_alcohol_consumption_per_day', 'reflux_history',
       'antireflux_treatment_types', 'h_pylori_infection',
       'initial_diagnosis_by', 'barretts_esophagus', 'goblet_cells_present',
       'history_of_esophageal_cancer', 'number_of_relatives_diagnosed',
```

```
       'has_new_tumor_events_information', 'day_of_form_completion',
       'month_of_form_completion', 'year_of_form_completion',
       'has_follow_ups_information', 'has_drugs_information',
       'has_radiations_information', 'project', 'stage_event_system_version',
       'stage_event_clinical_stage', 'stage_event_pathologic_stage',
       'stage_event_tnm_categories', 'stage_event_psa',
       'stage_event_gleason_grading', 'stage_event_ann_arbor',
       'stage_event_serum_markers', 'stage_event_igcccg_stage',
       'stage_event_masaoka_stage', 'primary_pathology_tumor_tissue_site',
       'primary_pathology_esophageal_tumor_cental_location',
       'primary_pathology_esophageal_tumor_involvement_sites',
       'primary_pathology_histological_type',
       'primary_pathology_columnar_metaplasia_present',
       'primary_pathology_columnar_mucosa_goblet_cell_present',
       'primary_pathology_columnar_mucosa_dysplasia',
       'primary_pathology_neoplasm_histologic_grade',
       'primary_pathology_days_to_initial_pathologic_diagnosis',
       'primary_pathology_age_at_initial_pathologic_diagnosis',
       'primary_pathology_year_of_initial_pathologic_diagnosis',
       'primary_pathology_initial_pathologic_diagnosis_method',
       'primary_pathology_init_pathology_dx_method_other',
       'primary_pathology_lymph_node_metastasis_radiographic_evidence',
       'primary_pathology_primary_lymph_node_presentation_assessment',
       'primary_pathology_lymph_node_examined_count',
       'primary_pathology_number_of_lymphnodes_positive_by_he',
       'primary_pathology_number_of_lymphnodes_positive_by_ihc',
       'primary_pathology_planned_surgery_status',
       'primary_pathology_treatment_prior_to_surgery',
       'primary_pathology_residual_tumor',
       'primary_pathology_karnofsky_performance_score',
       'primary_pathology_eastern_cancer_oncology_group',
       'primary_pathology_radiation_therapy',
       'primary_pathology_postoperative_rx_tx'],
      dtype='object')
```

[9]:
```
df = df.drop(['Unnamed: 0'], axis=1)
df.duplicated().sum()
```

[9]:
```
0
```

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

[10]:
```
patient_barcode                                 0
tissue_source_site                              0
patient_id                                      0
bcr_patient_uuid                                0
informed_consent_verified                       0
```

```
                                                    ...
primary_pathology_residual_tumor                     520
primary_pathology_karnofsky_performance_score       2625
primary_pathology_eastern_cancer_oncology_group     2628
primary_pathology_radiation_therapy                  638
primary_pathology_postoperative_rx_tx                658
Length: 84, dtype: int64
```

[11]:
```
null_percentage = (df.isnull().sum() / df.shape[0]) * 100

high_null_features = null_percentage[null_percentage>50]

high_null_features
```

[11]:
```
ethnicity                                                51.392723
days_to_death                                            69.962359
age_began_smoking_in_years                               56.537014
stopped_smoking_year                                     59.648683
antireflux_treatment_types                               74.981179
h_pylori_infection                                       60.928482
goblet_cells_present                                     89.485571
number_of_relatives_diagnosed                            78.946048
stage_event_clinical_stage                               66.925972
stage_event_psa                                         100.000000
stage_event_gleason_grading                             100.000000
stage_event_ann_arbor                                   100.000000
stage_event_serum_markers                               100.000000
stage_event_igcccg_stage                                100.000000
stage_event_masaoka_stage                               100.000000
primary_pathology_columnar_mucosa_goblet_cell_present    54.554580
primary_pathology_columnar_mucosa_dysplasia              56.085320
primary_pathology_init_pathology_dx_method_other         77.942284
primary_pathology_number_of_lymphnodes_positive_by_ihc   63.563363
primary_pathology_planned_surgery_status                 62.910916
primary_pathology_treatment_prior_to_surgery             71.442911
primary_pathology_karnofsky_performance_score            65.872020
primary_pathology_eastern_cancer_oncology_group          65.947302
dtype: float64
```

[12]:
```
features_to_drop = null_percentage[null_percentage>50].index

df = df.drop(columns = features_to_drop)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3985 entries, 0 to 3984
Data columns (total 61 columns):
```

```
 #   Column                                                  Non-Null
Count  Dtype
---  ------
--------------  -----
 0   patient_barcode                                         3985 non-
null   object
 1   tissue_source_site                                      3985 non-
null   object
 2   patient_id                                              3985 non-
null   object
 3   bcr_patient_uuid                                        3985 non-
null   object
 4   informed_consent_verified                               3985 non-
null   object
 5   icd_o_3_site                                            3985 non-
null   object
 6   icd_o_3_histology                                       3985 non-
null   object
 7   icd_10                                                  3985 non-
null   object
 8   tissue_prospective_collection_indicator                 3945 non-
null   object
 9   tissue_retrospective_collection_indicator               3945 non-
null   object
 10  days_to_birth                                           3985 non-
null   int64
 11  country_of_birth                                        2058 non-
null   object
 12  gender                                                  3985 non-
null   object
 13  height                                                  3766 non-
null   float64
 14  weight                                                  3945 non-
null   float64
 15  country_of_procurement                                  3945 non-
null   object
 16  state_province_of_procurement                           2705 non-
null   object
 17  city_of_procurement                                     3125 non-
null   object
 18  race_list                                               3566 non-
null   object
 19  other_dx                                                3985 non-
null   object
 20  history_of_neoadjuvant_treatment                        3985 non-
null   object
 21  person_neoplasm_cancer_status                           3650 non-
null   object
```

```
 22  vital_status                                          3985 non-
null   object
 23  days_to_last_followup                                 2788 non-
null   float64
 24  tobacco_smoking_history                               3605 non-
null   float64
 25  number_pack_years_smoked                              2169 non-
null   float64
 26  alcohol_history_documented                            3925 non-
null   object
 27  frequency_of_alcohol_consumption                      2469 non-
null   float64
 28  amount_of_alcohol_consumption_per_day                 2108 non-
null   float64
 29  reflux_history                                        3308 non-
null   object
 30  initial_diagnosis_by                                  3247 non-
null   object
 31  barretts_esophagus                                    3167 non-
null   object
 32  history_of_esophageal_cancer                          3148 non-
null   object
 33  has_new_tumor_events_information                      3985 non-
null   object
 34  day_of_form_completion                                3985 non-
null   int64
 35  month_of_form_completion                              3985 non-
null   int64
 36  year_of_form_completion                               3985 non-
null   int64
 37  has_follow_ups_information                            3985 non-
null   object
 38  has_drugs_information                                 3985 non-
null   object
 39  has_radiations_information                            3985 non-
null   object
 40  project                                               3985 non-
null   object
 41  stage_event_system_version                            3985 non-
null   object
 42  stage_event_pathologic_stage                          3487 non-
null   object
 43  stage_event_tnm_categories                            3985 non-
null   object
 44  primary_pathology_tumor_tissue_site                   3985 non-
null   object
 45  primary_pathology_esophageal_tumor_cental_location    3965 non-
null   object
```

```
 46  primary_pathology_esophageal_tumor_involvement_sites           3965 non-
null   object
 47  primary_pathology_histological_type                            3985 non-
null   object
 48  primary_pathology_columnar_metaplasia_present                  2390 non-
null   object
 49  primary_pathology_neoplasm_histologic_grade                    3985 non-
null   object
 50  primary_pathology_days_to_initial_pathologic_diagnosis         3985 non-
null   int64
 51  primary_pathology_age_at_initial_pathologic_diagnosis          3985 non-
null   int64
 52  primary_pathology_year_of_initial_pathologic_diagnosis         3845 non-
null   float64
 53  primary_pathology_initial_pathologic_diagnosis_method          3885 non-
null   object
 54  primary_pathology_lymph_node_metastasis_radiographic_evidence  3148 non-
null   object
 55  primary_pathology_primary_lymph_node_presentation_assessment   3665 non-
null   object
 56  primary_pathology_lymph_node_examined_count                    2985 non-
null   float64
 57  primary_pathology_number_of_lymphnodes_positive_by_he          2985 non-
null   float64
 58  primary_pathology_residual_tumor                               3465 non-
null   object
 59  primary_pathology_radiation_therapy                            3347 non-
null   object
 60  primary_pathology_postoperative_rx_tx                          3327 non-
null   object
dtypes: float64(10), int64(6), object(45)
memory usage: 1.9+ MB
```

[13]:
```
df.describe()
```

[13]:

	days_to_birth	height	weight	days_to_last_followup	\
count	3985.000000	3766.000000	3945.000000	2788.000000	
mean	-23367.341782	172.128518	75.622560	306.201937	
std	4441.493885	9.080075	18.997044	506.175392	
min	-32972.000000	145.000000	41.000000	-4.000000	
25%	-27075.000000	166.000000	62.000000	3.000000	
50%	-22812.000000	173.000000	72.000000	105.000000	
75%	-19925.000000	178.000000	86.000000	408.000000	
max	-10143.000000	202.000000	198.000000	3714.000000	

	tobacco_smoking_history	number_pack_years_smoked	\
count	3605.000000	2169.000000	

```
mean                  2.362829                 35.392577
std                   1.142633                 21.614376
min                   1.000000                  1.000000
25%                   1.000000                 19.000000
50%                   2.000000                 31.000000
75%                   3.000000                 50.000000
max                   4.000000                102.000000

       frequency_of_alcohol_consumption  \
count                       2469.000000
mean                           3.523289
std                            3.130464
min                            0.000000
25%                            0.000000
50%                            3.000000
75%                            7.000000
max                            7.000000

       amount_of_alcohol_consumption_per_day  day_of_form_completion  \
count                            2108.000000             3985.000000
mean                                1.749051               16.468758
std                                 2.227695                8.123982
min                                 0.000000                1.000000
25%                                 0.000000               11.000000
50%                                 1.000000               16.000000
75%                                 2.000000               25.000000
max                                14.000000               30.000000

       month_of_form_completion  year_of_form_completion  \
count               3985.000000              3985.000000
mean                   4.812547              2013.545546
std                    3.743568                 0.598690
min                    1.000000              2012.000000
25%                    2.000000              2013.000000
50%                    3.000000              2014.000000
75%                    8.000000              2014.000000
max                   12.000000              2015.000000

       primary_pathology_days_to_initial_pathologic_diagnosis  \
count                                             3985.0
mean                                                 0.0
std                                                  0.0
min                                                  0.0
25%                                                  0.0
50%                                                  0.0
75%                                                  0.0
max                                                  0.0
```

```
       primary_pathology_age_at_initial_pathologic_diagnosis  \
count                                          3985.000000
mean                                             63.480050
std                                              12.182604
min                                              27.000000
25%                                              54.000000
50%                                              62.000000
75%                                              74.000000
max                                              90.000000

       primary_pathology_year_of_initial_pathologic_diagnosis  \
count                                          3845.000000
mean                                           2009.237451
std                                               4.204706
min                                            1998.000000
25%                                            2007.000000
50%                                            2011.000000
75%                                            2012.000000
max                                            2013.000000

       primary_pathology_lymph_node_examined_count  \
count                                  2985.000000
mean                                     14.269347
std                                      12.187865
min                                       1.000000
25%                                       5.000000
50%                                      12.000000
75%                                      19.000000
max                                      87.000000

       primary_pathology_number_of_lymphnodes_positive_by_he
count                                          2985.000000
mean                                              2.450251
std                                               3.324540
min                                               0.000000
25%                                               0.000000
50%                                               1.000000
75%                                               4.000000
max                                              21.000000
```

[14]:
```
df.nunique()
```

[14]:
```
patient_barcode                                    3985
tissue_source_site                                   19
patient_id                                          185
bcr_patient_uuid                                    185
```

```
informed_consent_verified                                1
                                                        ...
primary_pathology_lymph_node_examined_count             39
primary_pathology_number_of_lymphnodes_positive_by_he   14
primary_pathology_residual_tumor                         4
primary_pathology_radiation_therapy                      2
primary_pathology_postoperative_rx_tx                    2
Length: 61, dtype: int64
```

[16]:
```
object_columns = df.select_dtypes(include=['object']).columns
print("Object type columns:")
print(object_columns)
```

```
Object type columns:
Index(['patient_barcode', 'tissue_source_site', 'patient_id',
       'bcr_patient_uuid', 'informed_consent_verified', 'icd_o_3_site',
       'icd_o_3_histology', 'icd_10',
       'tissue_prospective_collection_indicator',
       'tissue_retrospective_collection_indicator', 'country_of_birth',
       'gender', 'country_of_procurement', 'state_province_of_procurement',
       'city_of_procurement', 'race_list', 'other_dx',
       'history_of_neoadjuvant_treatment', 'person_neoplasm_cancer_status',
       'vital_status', 'alcohol_history_documented', 'reflux_history',
       'initial_diagnosis_by', 'barretts_esophagus',
       'history_of_esophageal_cancer', 'has_new_tumor_events_information',
       'has_follow_ups_information', 'has_drugs_information',
       'has_radiations_information', 'project', 'stage_event_system_version',
       'stage_event_pathologic_stage', 'stage_event_tnm_categories',
       'primary_pathology_tumor_tissue_site',
       'primary_pathology_esophageal_tumor_cental_location',
       'primary_pathology_esophageal_tumor_involvement_sites',
       'primary_pathology_histological_type',
       'primary_pathology_columnar_metaplasia_present',
       'primary_pathology_neoplasm_histologic_grade',
       'primary_pathology_initial_pathologic_diagnosis_method',
       'primary_pathology_lymph_node_metastasis_radiographic_evidence',
       'primary_pathology_primary_lymph_node_presentation_assessment',
       'primary_pathology_residual_tumor',
       'primary_pathology_radiation_therapy',
       'primary_pathology_postoperative_rx_tx'],
      dtype='object')
```

[17]:
```
numerical_columns = df.select_dtypes(include=['int64','float64']).columns
print("\n Numerical type columns:")
print(numerical_columns)
```

```
 Numerical type columns:
```

```
Index(['days_to_birth', 'height', 'weight', 'days_to_last_followup',
       'tobacco_smoking_history', 'number_pack_years_smoked',
       'frequency_of_alcohol_consumption',
       'amount_of_alcohol_consumption_per_day', 'day_of_form_completion',
       'month_of_form_completion', 'year_of_form_completion',
       'primary_pathology_days_to_initial_pathologic_diagnosis',
       'primary_pathology_age_at_initial_pathologic_diagnosis',
       'primary_pathology_year_of_initial_pathologic_diagnosis',
       'primary_pathology_lymph_node_examined_count',
       'primary_pathology_number_of_lymphnodes_positive_by_he'],
      dtype='object')
```

[18]:
```python
def calssify_features(df):
    categorical_features = []
    non_categorical_features = []
    discrete_features = []
    continuous_features = []

    for column in df.columns:
        if df[column].dtype == 'object':
            if df[column].nunique() < 10:
                categorical_features.append(column)
            else:
                non_categorical_features.append(column)
        elif df[column].dtype in ['int64', 'float64']:
            if df[column].nunique() < 10:
                discrete_features.append(column)
            else:
                continuous_features.append(column)

    return categorical_features, non_categorical_features, discrete_features,␣
 ↪continuous_features


categorical_features, non_categorical_features, discrete_features,␣
 ↪continuous_features = calssify_features(df)

print("Categorical features:", categorical_features)
print("Non-categorical features:", non_categorical_features)

print("Discrete features:", discrete_features)

print("Continuous features:", continuous_features)
```

```
Categorical features: ['informed_consent_verified', 'icd_o_3_site',
'icd_o_3_histology', 'icd_10', 'tissue_prospective_collection_indicator',
'tissue_retrospective_collection_indicator', 'country_of_birth', 'gender',
'race_list', 'other_dx', 'history_of_neoadjuvant_treatment',
```

```
'person_neoplasm_cancer_status', 'vital_status', 'alcohol_history_documented',
'reflux_history', 'initial_diagnosis_by', 'barretts_esophagus',
'history_of_esophageal_cancer', 'has_new_tumor_events_information',
'has_follow_ups_information', 'has_drugs_information',
'has_radiations_information', 'project', 'stage_event_system_version',
'primary_pathology_tumor_tissue_site',
'primary_pathology_esophageal_tumor_cental_location',
'primary_pathology_esophageal_tumor_involvement_sites',
'primary_pathology_histological_type',
'primary_pathology_columnar_metaplasia_present',
'primary_pathology_neoplasm_histologic_grade',
'primary_pathology_initial_pathologic_diagnosis_method',
'primary_pathology_lymph_node_metastasis_radiographic_evidence',
'primary_pathology_primary_lymph_node_presentation_assessment',
'primary_pathology_residual_tumor', 'primary_pathology_radiation_therapy',
'primary_pathology_postoperative_rx_tx']
Non-categorical features: ['patient_barcode', 'tissue_source_site',
'patient_id', 'bcr_patient_uuid', 'country_of_procurement',
'state_province_of_procurement', 'city_of_procurement',
'stage_event_pathologic_stage', 'stage_event_tnm_categories']
Discrete features: ['tobacco_smoking_history',
'frequency_of_alcohol_consumption', 'year_of_form_completion',
'primary_pathology_days_to_initial_pathologic_diagnosis']
Continuous features: ['days_to_birth', 'height', 'weight',
'days_to_last_followup', 'number_pack_years_smoked',
'amount_of_alcohol_consumption_per_day', 'day_of_form_completion',
'month_of_form_completion',
'primary_pathology_age_at_initial_pathologic_diagnosis',
'primary_pathology_year_of_initial_pathologic_diagnosis',
'primary_pathology_lymph_node_examined_count',
'primary_pathology_number_of_lymphnodes_positive_by_he']
```

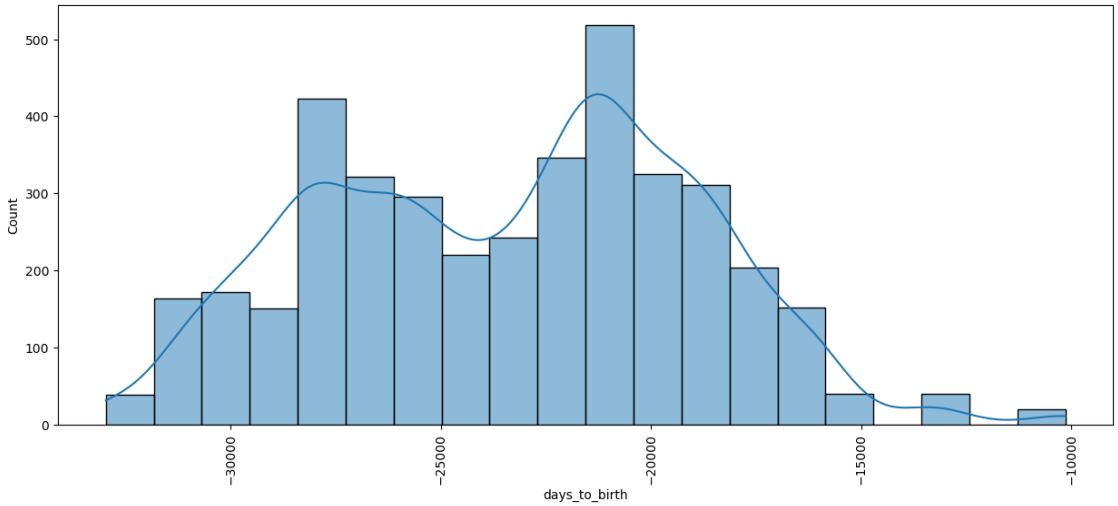
[21]:
```
df[categorical_features]=df[categorical_features].fillna("Not Available")

df[non_categorical_features] = df[non_categorical_features].fillna("Not␣
 ↪Available")

for features in discrete_features:
    mode_value = df[features].mode()[0]
    df[features] = df[features].fillna(mode_value)

for features in continuous_features:
    mean_value = df[features].mean()
    df[features] = df[features].fillna(mean_value)
```

[22]:
```
df.isnull().sum()
```

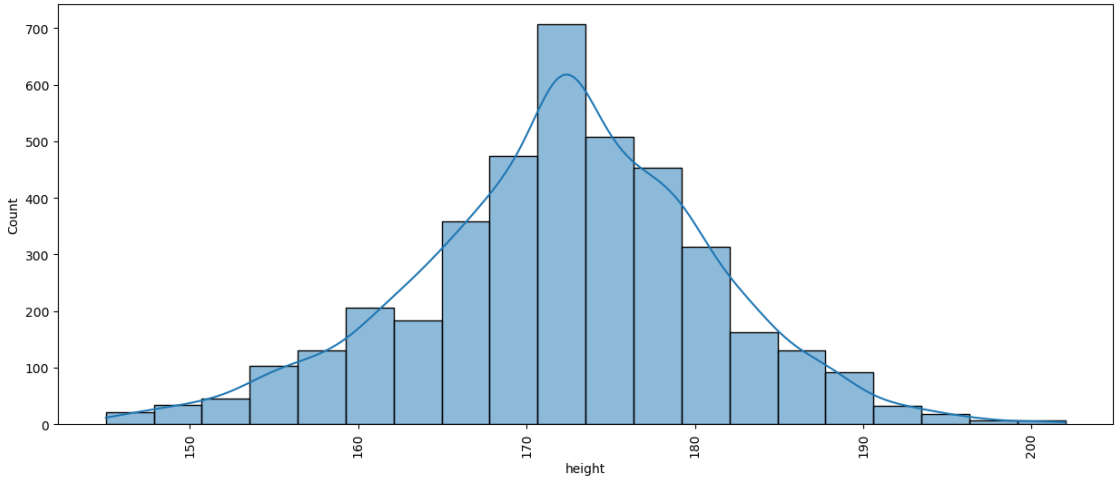
[22]:
```
patient_barcode                                          0
tissue_source_site                                       0
patient_id                                               0
bcr_patient_uuid                                         0
informed_consent_verified                                0
                                                        ..
primary_pathology_lymph_node_examined_count              0
primary_pathology_number_of_lymphnodes_positive_by_he    0
primary_pathology_residual_tumor                         0
primary_pathology_radiation_therapy                      0
primary_pathology_postoperative_rx_tx                    0
Length: 61, dtype: int64
```

[23]:
```
for i in continuous_features:
    plt.figure(figsize=(15,6))
    sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
    plt.xticks(rotation = 90)
    plt.show()
```

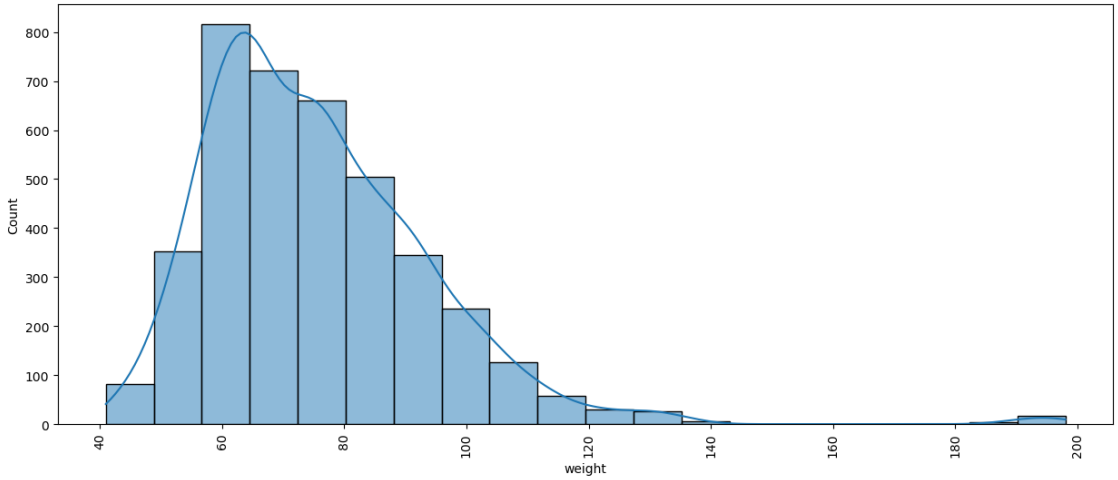
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



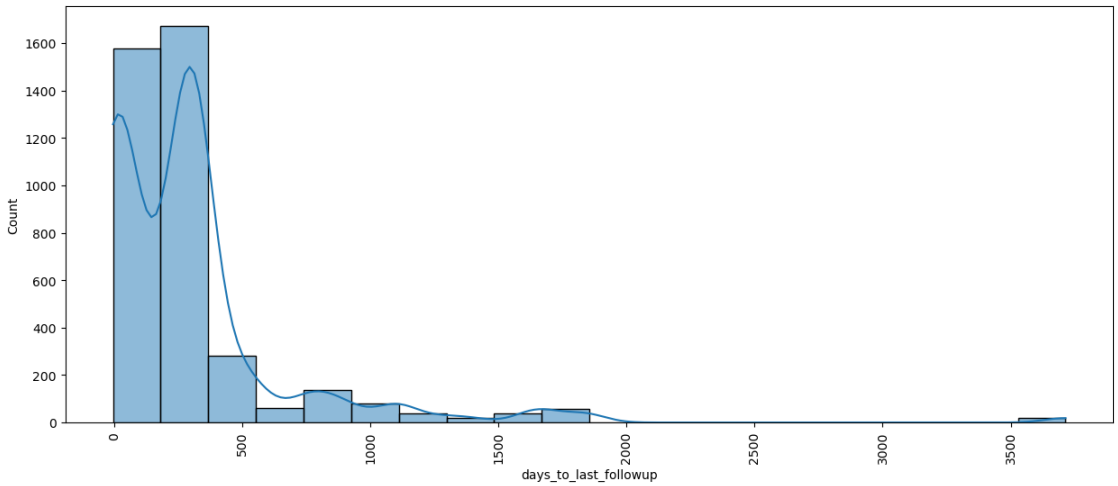

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



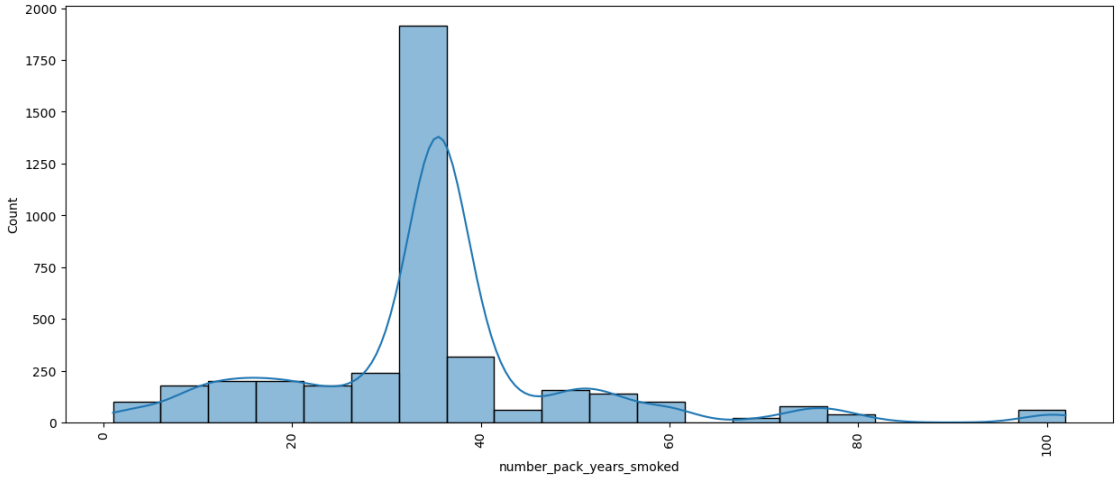

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



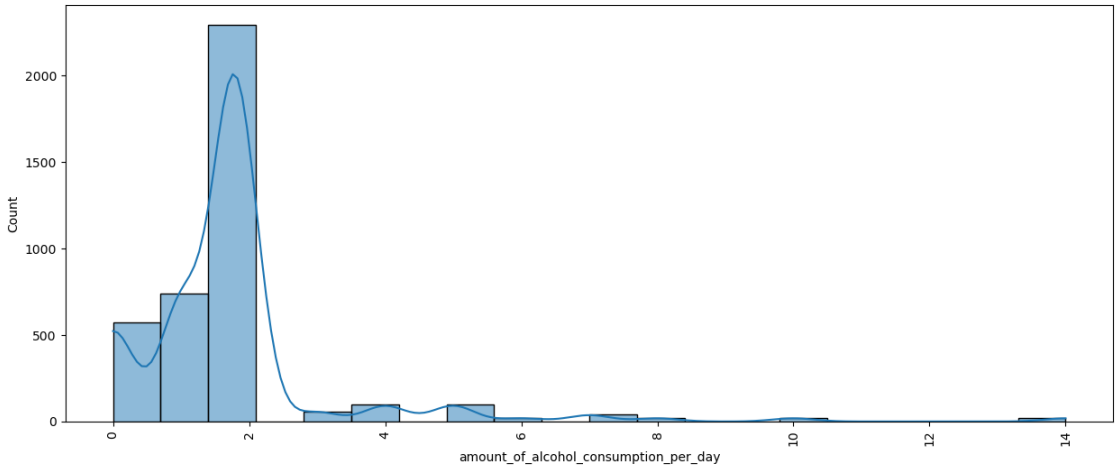

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



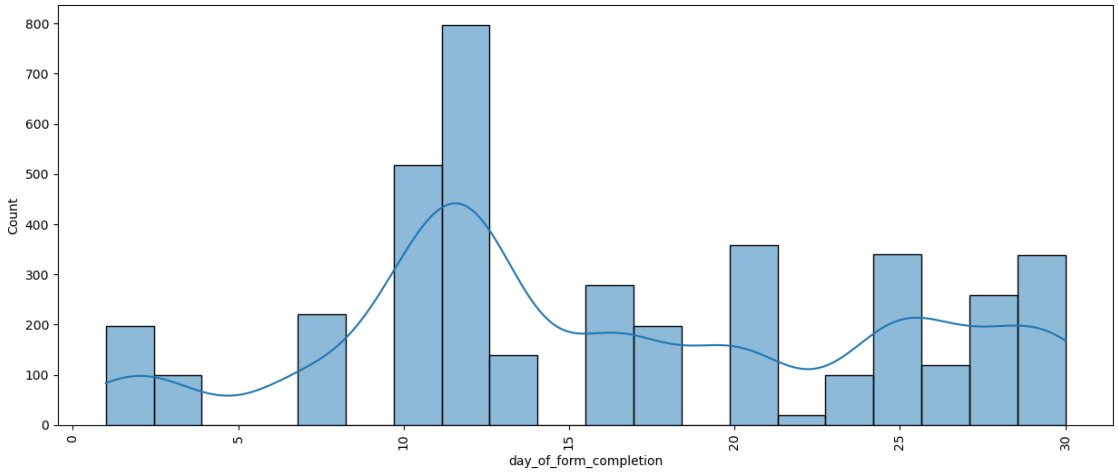

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



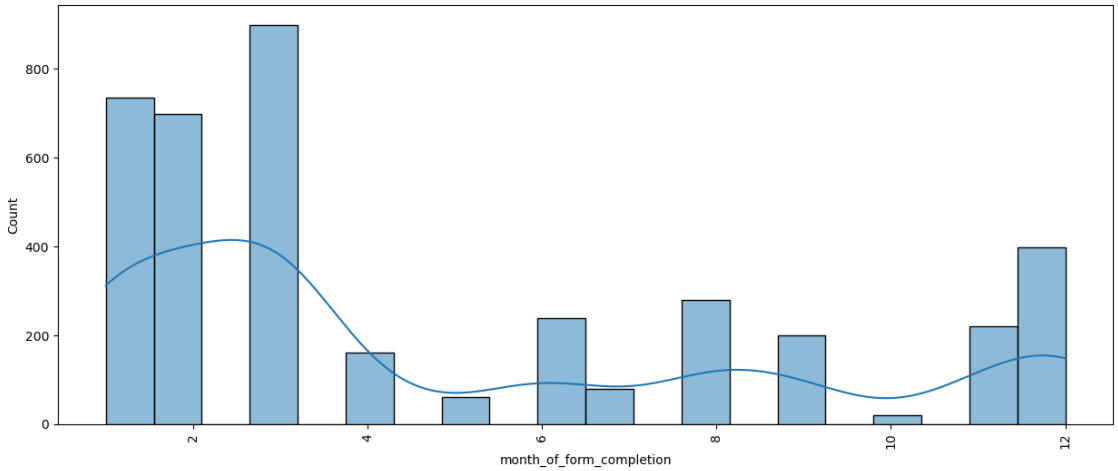

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



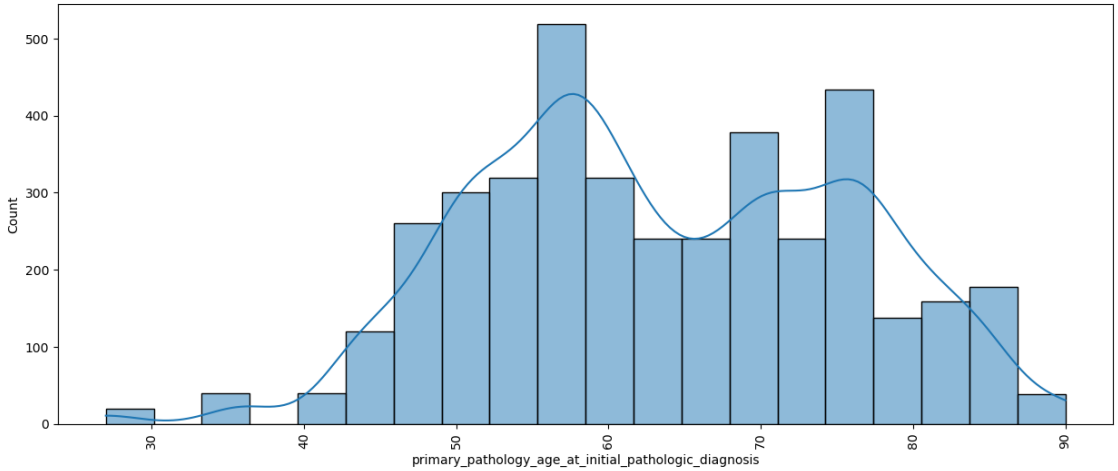

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



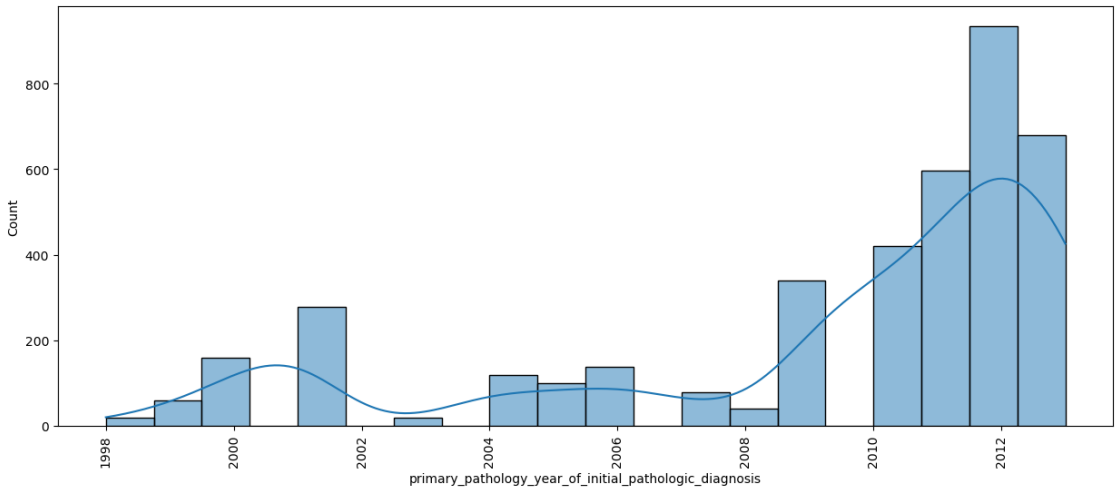

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



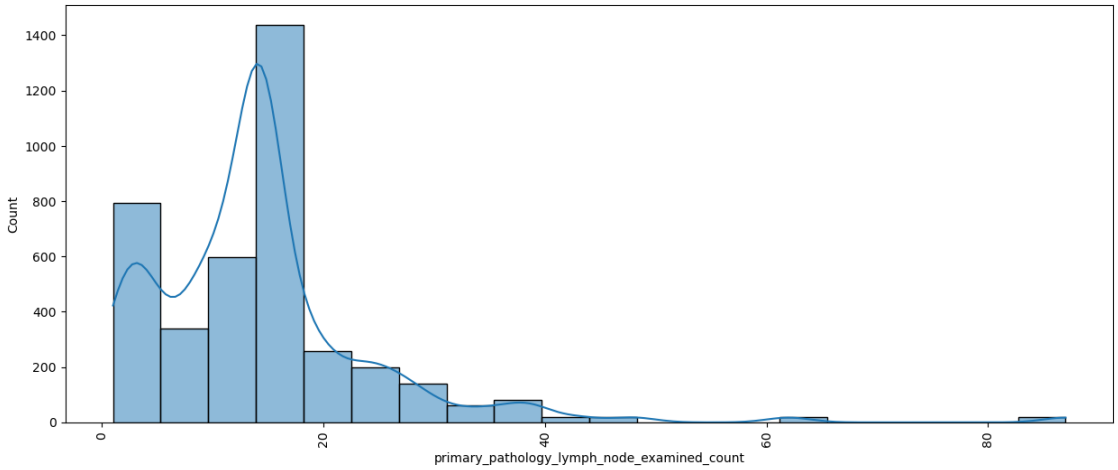

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```



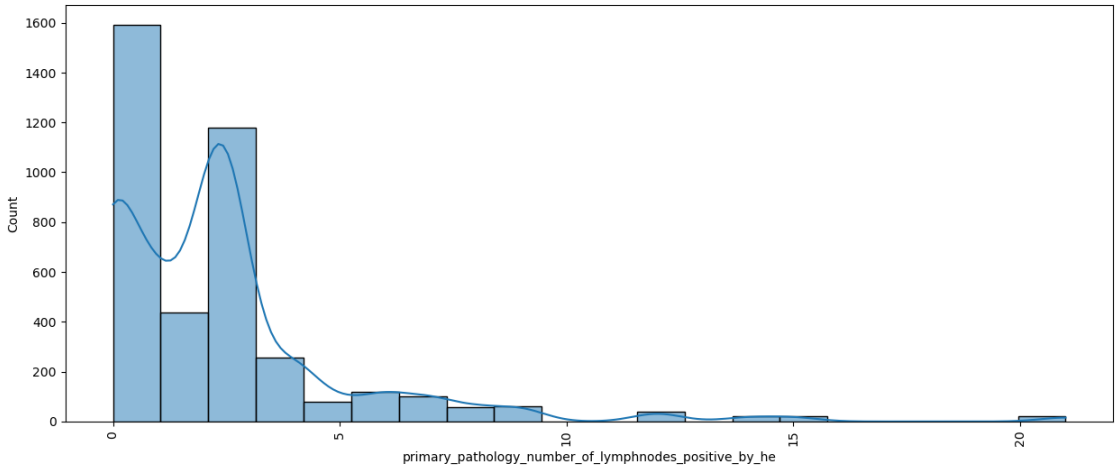

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```

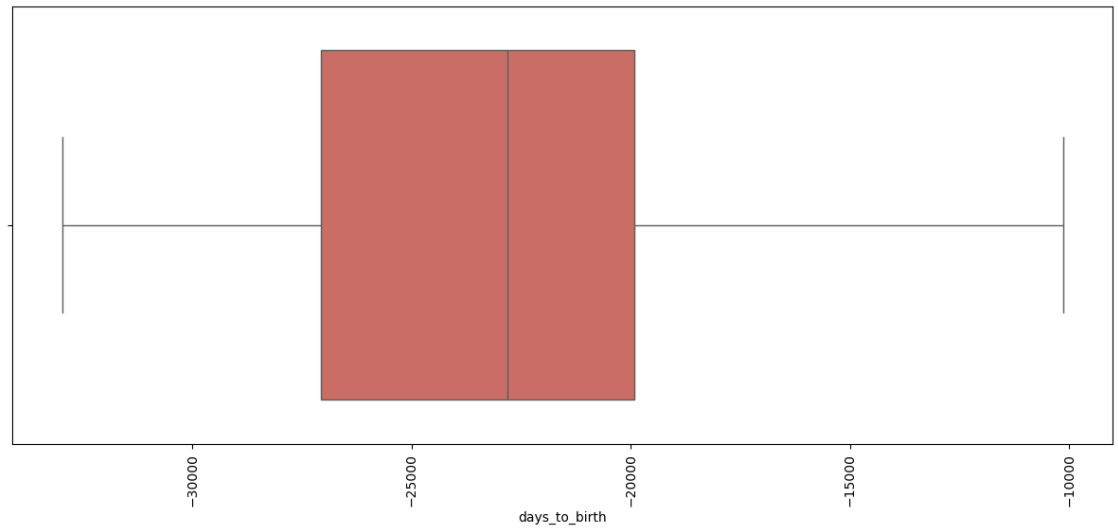



```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\3590757880.py:3: UserWarning:
Ignoring `palette` because no `hue` variable has been assigned.
  sns.histplot(df[i], bins = 20, kde = True, palette = 'hls')
```

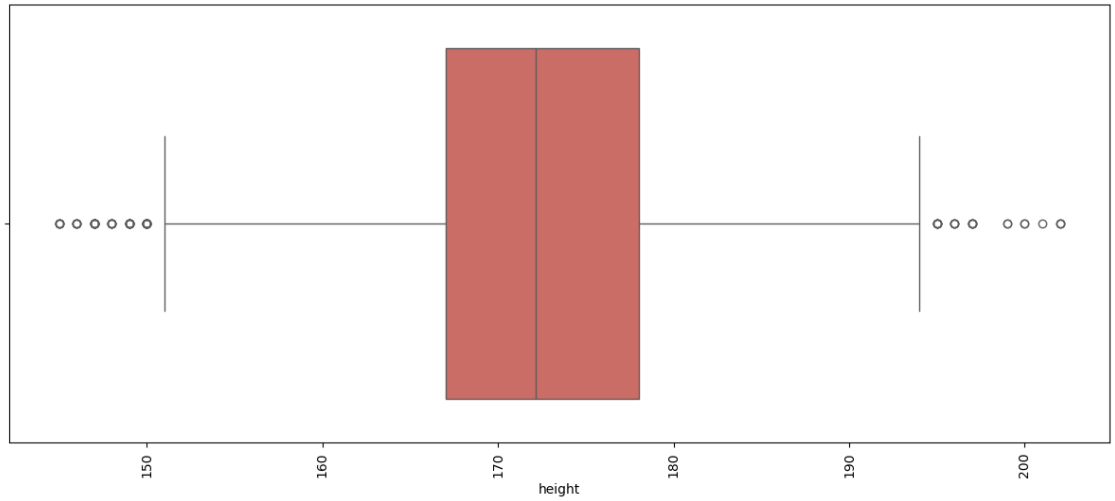


[26]:
```
for i in continuous_features:
    plt.figure(figsize=(15,6))
    sns.boxplot(x=i, data=df, palette='hls')
    plt.xticks(rotation = 90)
    plt.show()
```

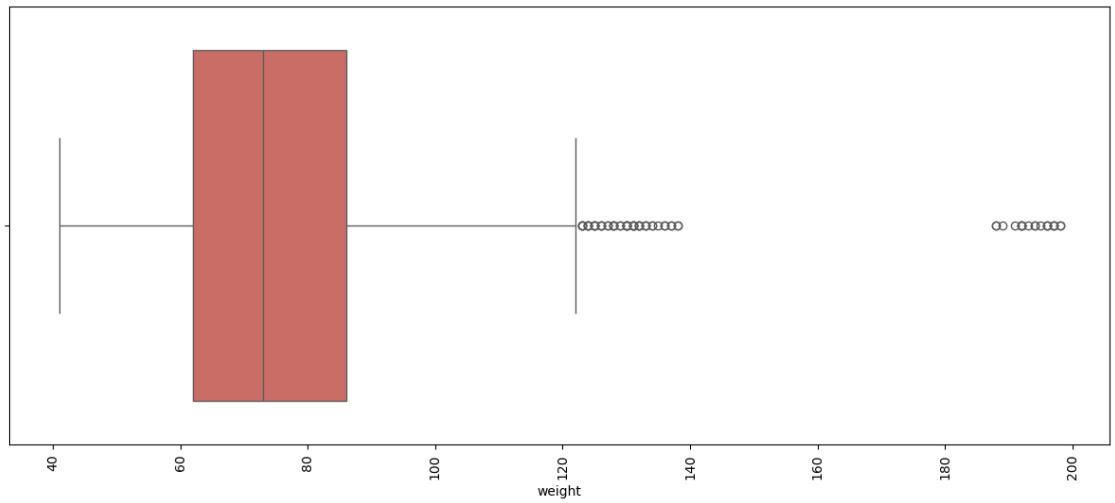
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```



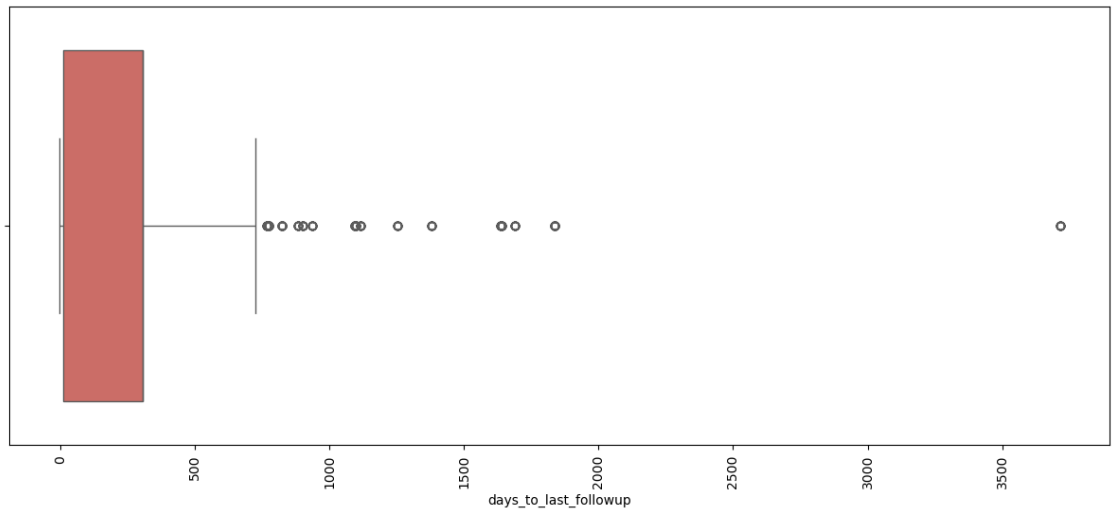
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```



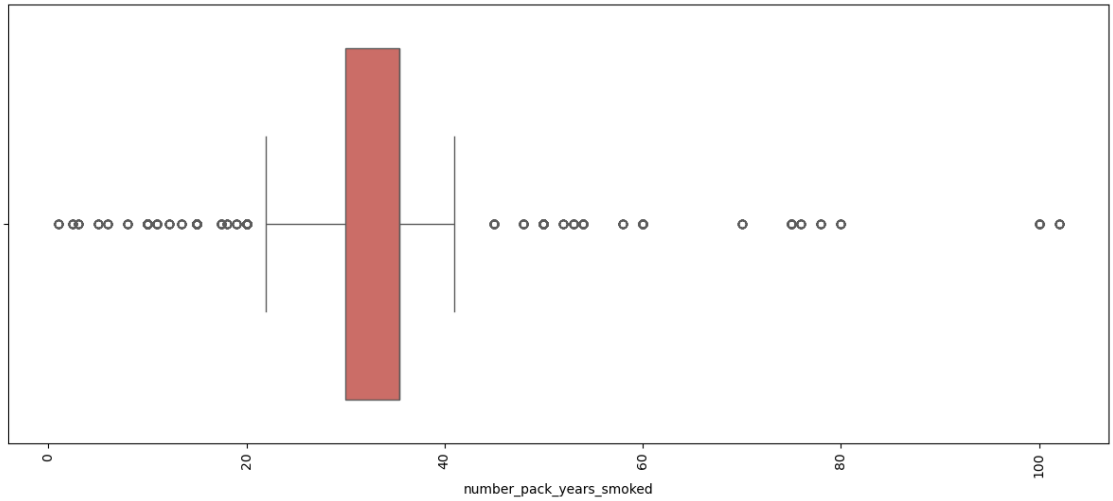

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```



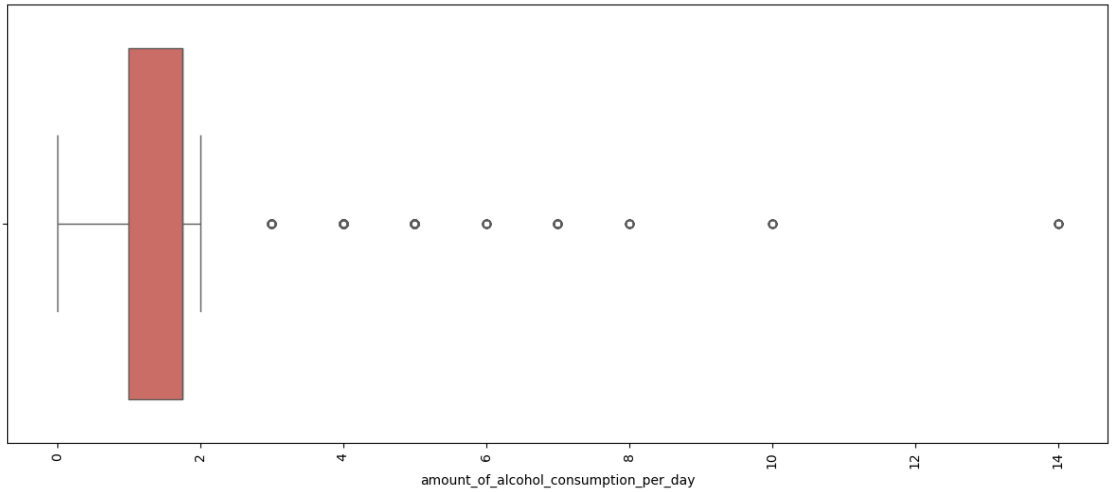
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```

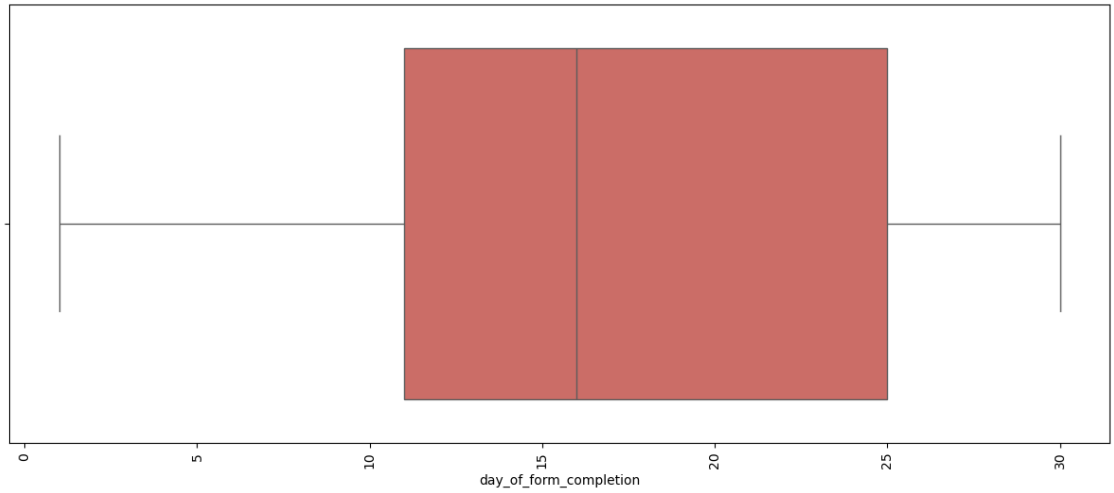



```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```

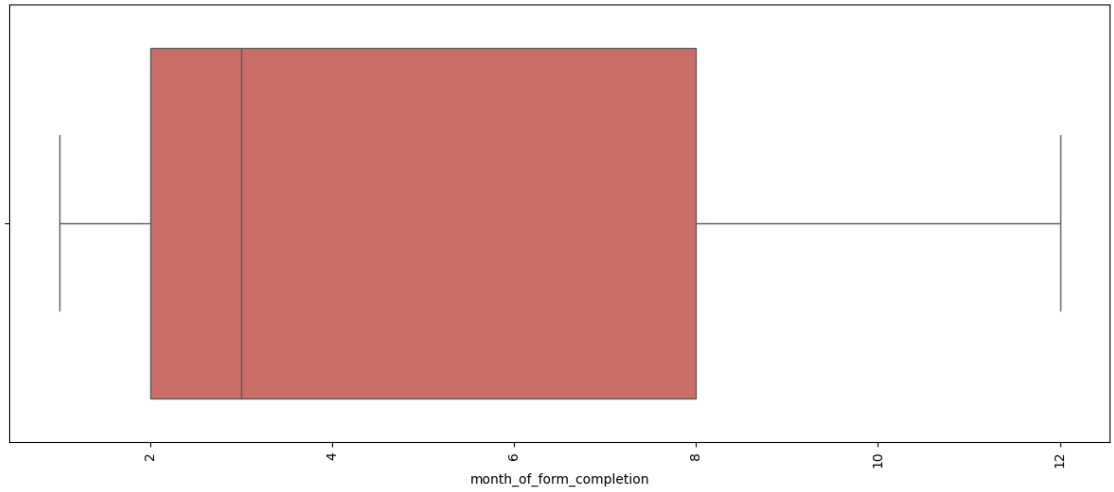



```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

```
  sns.boxplot(x=i, data=df, palette='hls')
```



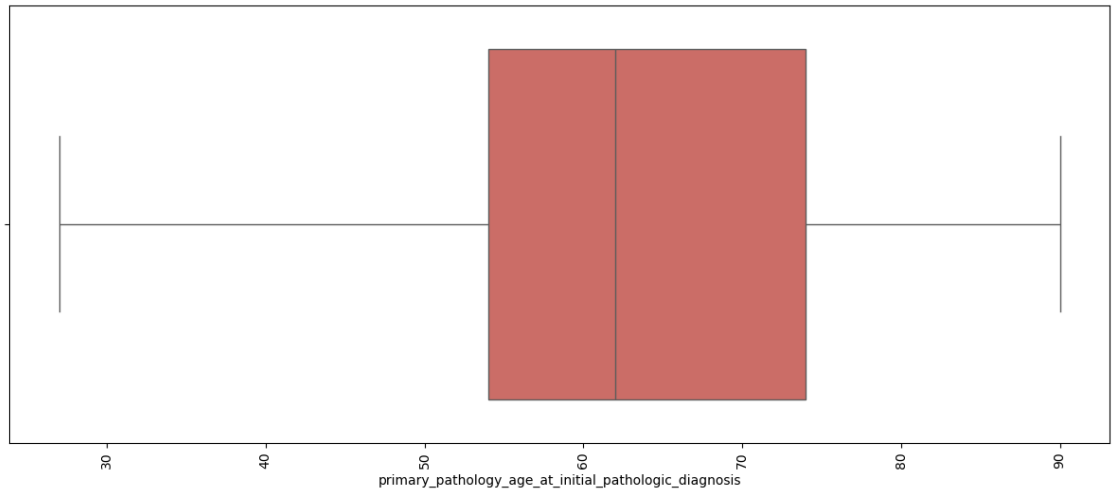

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```



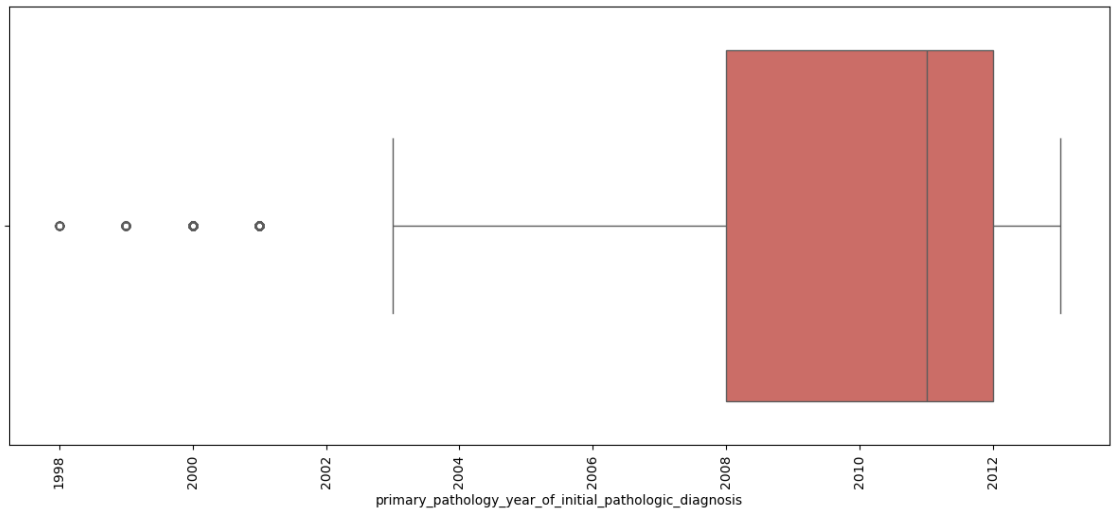

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

effect.

```
  sns.boxplot(x=i, data=df, palette='hls')
```

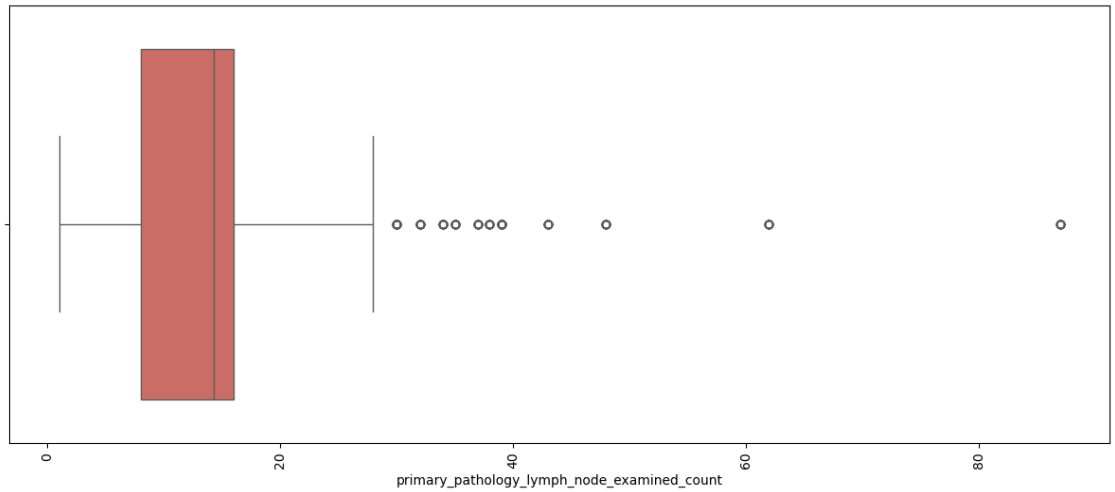



```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```

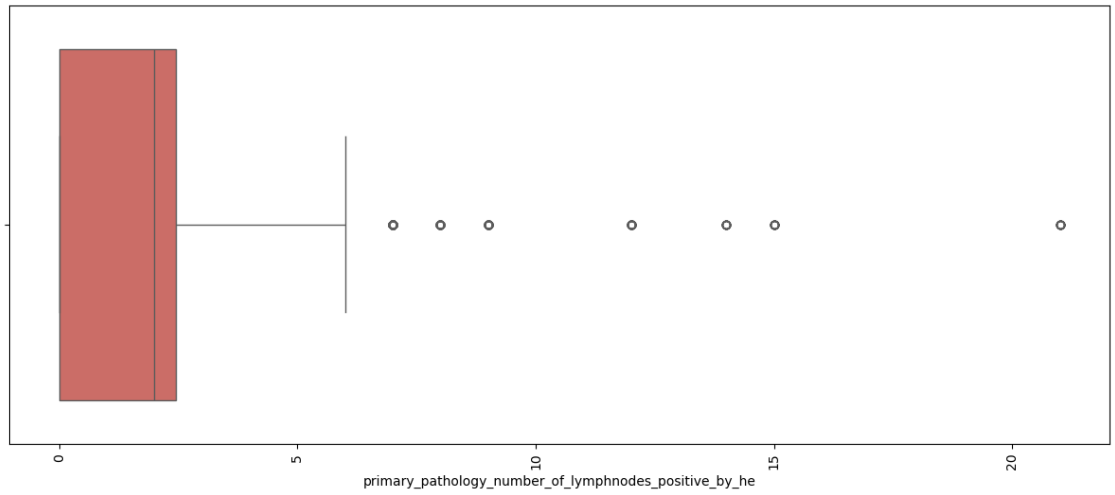



```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:
```

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

  sns.boxplot(x=i, data=df, palette='hls')
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1239036499.py:3: FutureWarning:

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

  sns.boxplot(x=i, data=df, palette='hls')
```



[27]:
```
for i in discrete_features:
    print(df[i].unique())
    print()
```

```
[1. 2. 3. 4.]

[7. 1. 2. 5. 3. 4. 0.]

[2014 2012 2013 2015]

[0]
```

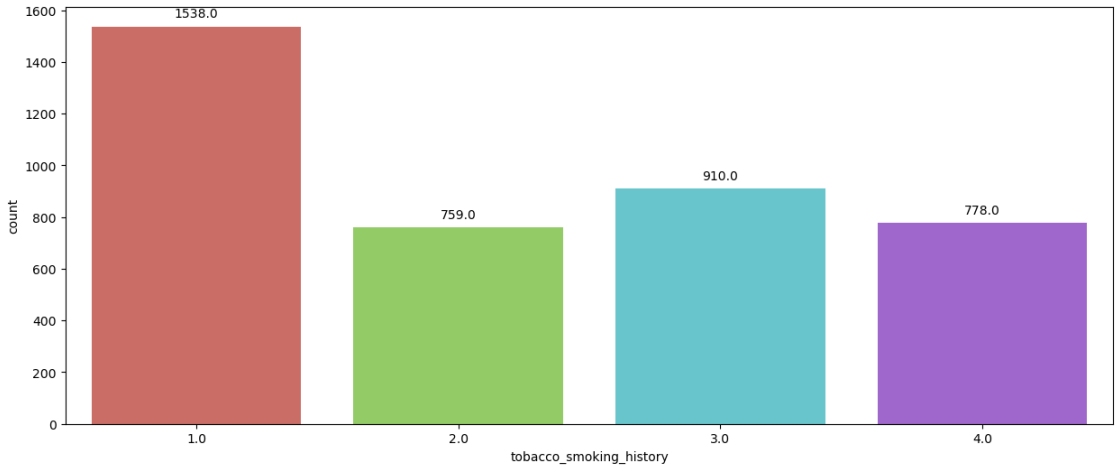
[28]:
```
for i in discrete_features:
    print(i)
    print(df[i].value_counts())
    print()
```

```
tobacco_smoking_history
tobacco_smoking_history
1.0    1538
3.0     910
4.0     778
2.0     759
Name: count, dtype: int64

frequency_of_alcohol_consumption
frequency_of_alcohol_consumption
7.0    2514
0.0     794
1.0     258
5.0     139
2.0     120
3.0     120
4.0      40
Name: count, dtype: int64

year_of_form_completion
year_of_form_completion
2014    2094
2013    1671
2012     120
2015     100
Name: count, dtype: int64

primary_pathology_days_to_initial_pathologic_diagnosis
primary_pathology_days_to_initial_pathologic_diagnosis
0    3985
```

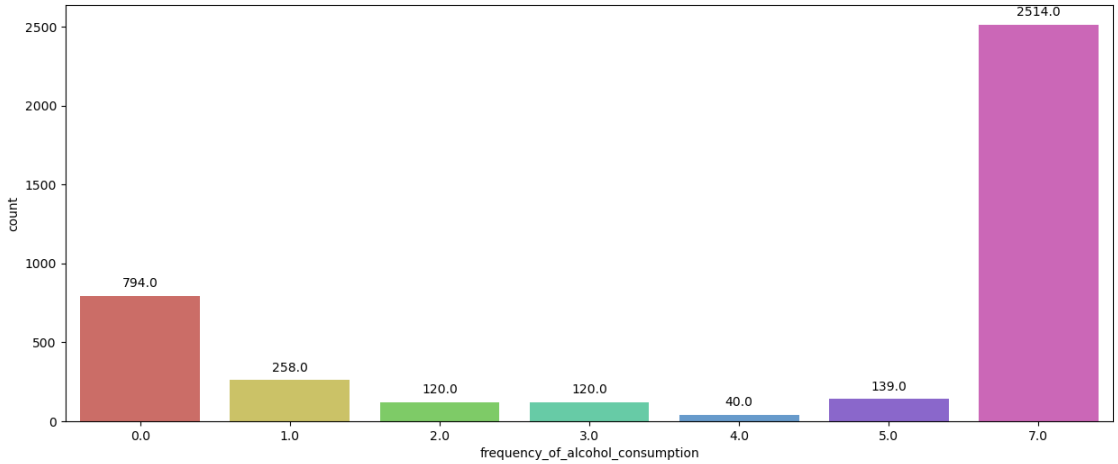
```
Name: count, dtype: int64
```

[30]:
```python
for i in discrete_features:
    plt.figure(figsize=(15,6))
    ax = sns.countplot(x=i, data=df, palette = 'hls')

    for p in ax.patches:
        height = p.get_height()
        ax.annotate(f'{height}',
                    xy=(p.get_x() + p.get_width() / 2, height),
                    xytext = (0, 10),
                    textcoords = 'offset points',
                    ha = 'center', va = 'center')

    plt.show()
```

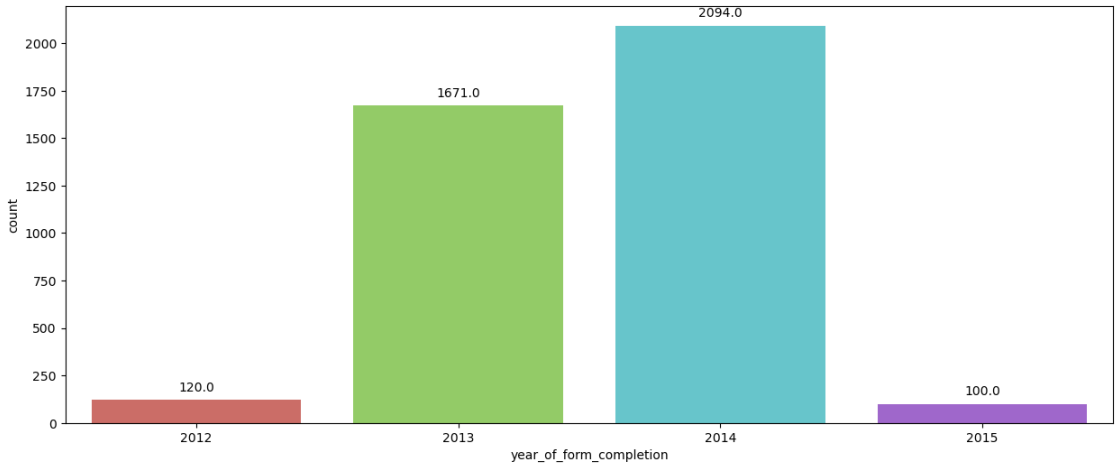
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1876973769.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



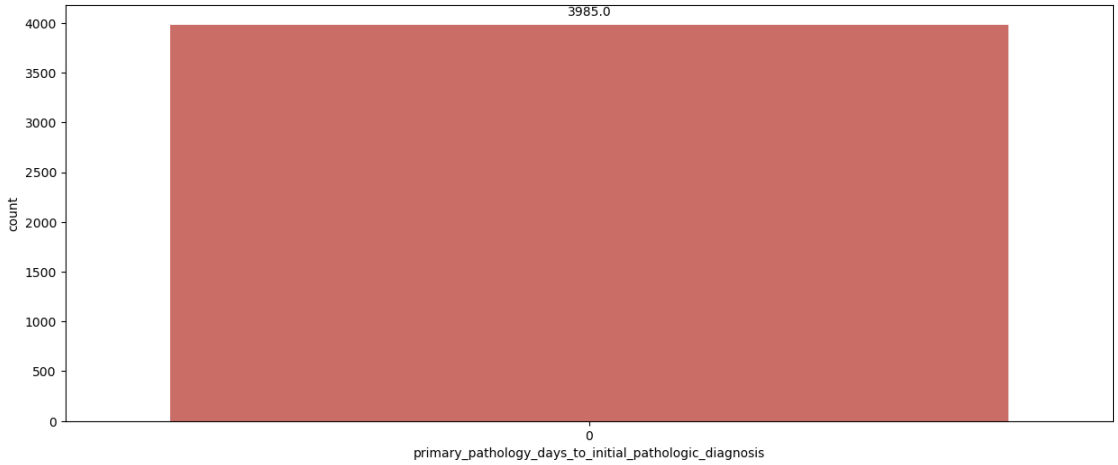

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1876973769.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1876973769.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1876973769.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



[36]:
```
import plotly.express as px

for i in discrete_features:
    counts = df[i].value_counts()
    fig = px.pie(counts, values = counts.values, names = counts.index, title =␣
 ↪f'Distribution of {i}')
    fig.show()

for i in categorical_features:
    print(i)
    print(df[i].unique())
    print()
```

```
informed_consent_verified
['YES']

icd_o_3_site
['C15.5' 'C15.9' 'C15.4' 'C15.1' 'C15.3' 'C16.0']

icd_o_3_histology
['8140/3' '8070/3' '8071/3' '8480/3' '8083/3' '8211/3']

icd_10
['C15.5' 'C15.9' 'C15.4' 'C15.3' 'C16.0']

tissue_prospective_collection_indicator
['NO' 'YES' 'Not Available']

tissue_retrospective_collection_indicator
['YES' 'NO' 'Not Available']
```

```
country_of_birth
['Not Available' 'Russia' 'Ukraine' 'Vietnam' 'Bulgaria' 'United States'
 'Australia' 'Brazil' 'United Kingdom']

gender
['MALE' 'FEMALE']

race_list
['Not Available' 'WHITE' 'ASIAN' 'BLACK OR AFRICAN AMERICAN']

other_dx
['No' 'Yes']

history_of_neoadjuvant_treatment
['No']

person_neoplasm_cancer_status
['WITH TUMOR' 'TUMOR FREE' 'Not Available']

vital_status
['Dead' 'Alive']

alcohol_history_documented
['NO' 'YES' 'Not Available']

reflux_history
['Not Available' 'NO' 'YES']

initial_diagnosis_by
['Symptomatic' 'Not Available' 'Screening' 'Surveillance']

barretts_esophagus
['No' 'Yes-UK' 'Yes-USA' 'Not Available']

history_of_esophageal_cancer
['Not Available' 'NO' 'YES']

has_new_tumor_events_information
['YES' 'NO']

has_follow_ups_information
['NO' 'YES']

has_drugs_information
['NO' 'YES']

has_radiations_information
['NO' 'YES']
```

```
project
['TCGA-ESCA']

stage_event_system_version
['5th' '7th' '6th']

primary_pathology_tumor_tissue_site
['Esophagus']

primary_pathology_esophageal_tumor_cental_location
['Distal' 'Mid' 'Proximal' 'Not Available']

primary_pathology_esophageal_tumor_involvement_sites
['Distal' 'Mid' 'Proximal' 'MidDistal' 'ProximalMid' 'Not Available']

primary_pathology_histological_type
['Esophagus Adenocarcinoma, NOS' 'Esophagus Squamous Cell Carcinoma']

primary_pathology_columnar_metaplasia_present
['NO' 'YES' 'Not Available']

primary_pathology_neoplasm_histologic_grade
['G3' 'G2' 'GX' 'G1']

primary_pathology_initial_pathologic_diagnosis_method
['Other method, specify:' 'Endoscopic Biopsy' 'Surgical Resection'
 'Not Available']

primary_pathology_lymph_node_metastasis_radiographic_evidence
['YES' 'NO' 'Not Available']

primary_pathology_primary_lymph_node_presentation_assessment
['YES' 'Not Available' 'NO']

primary_pathology_residual_tumor
['R1' 'R0' 'RX' 'Not Available' 'R2']

primary_pathology_radiation_therapy
['NO' 'Not Available' 'YES']

primary_pathology_postoperative_rx_tx
['NO' 'YES' 'Not Available']
```

[38]:
```
for i in categorical_features:
    print(i)
```

```
    print(df[i].value_counts())
    print()
```

```
informed_consent_verified
informed_consent_verified
YES    3985
Name: count, dtype: int64

icd_o_3_site
icd_o_3_site
C15.5    2727
C15.4     859
C15.9     200
C15.3     100
C16.0      59
C15.1      40
Name: count, dtype: int64

icd_o_3_histology
icd_o_3_histology
8140/3    1968
8070/3    1857
8071/3     100
8480/3      20
8083/3      20
8211/3      20
Name: count, dtype: int64

icd_10
icd_10
C15.5    2727
C15.4     859
C15.9     240
C15.3     100
C16.0      59
Name: count, dtype: int64

tissue_prospective_collection_indicator
tissue_prospective_collection_indicator
NO               2368
YES              1577
Not Available      40
Name: count, dtype: int64

tissue_retrospective_collection_indicator
tissue_retrospective_collection_indicator
YES              2368
NO               1577
```

```
Not Available      40
Name: count, dtype: int64

country_of_birth
country_of_birth
Not Available     1927
Vietnam            840
United States      418
Brazil             380
Russia             240
Ukraine            120
Bulgaria            20
Australia           20
United Kingdom      20
Name: count, dtype: int64

gender
gender
MALE      3369
FEMALE     616
Name: count, dtype: int64

race_list
race_list
WHITE                         2546
ASIAN                          920
Not Available                  419
BLACK OR AFRICAN AMERICAN      100
Name: count, dtype: int64

other_dx
other_dx
No     3439
Yes     546
Name: count, dtype: int64

history_of_neoadjuvant_treatment
history_of_neoadjuvant_treatment
No    3985
Name: count, dtype: int64

person_neoplasm_cancer_status
person_neoplasm_cancer_status
TUMOR FREE       2235
WITH TUMOR       1415
Not Available     335
Name: count, dtype: int64
```

```
vital_status
vital_status
Alive    2788
Dead     1197
Name: count, dtype: int64

alcohol_history_documented
alcohol_history_documented
YES              2846
NO               1079
Not Available      60
Name: count, dtype: int64

reflux_history
reflux_history
NO               2073
YES              1235
Not Available     677
Name: count, dtype: int64

initial_diagnosis_by
initial_diagnosis_by
Symptomatic      2930
Not Available     738
Screening         257
Surveillance       60
Name: count, dtype: int64

barretts_esophagus
barretts_esophagus
No               2569
Not Available     818
Yes-USA           398
Yes-UK            200
Name: count, dtype: int64

history_of_esophageal_cancer
history_of_esophageal_cancer
NO               2889
Not Available     837
YES               259
Name: count, dtype: int64

has_new_tumor_events_information
has_new_tumor_events_information
NO     2432
YES    1553
Name: count, dtype: int64
```

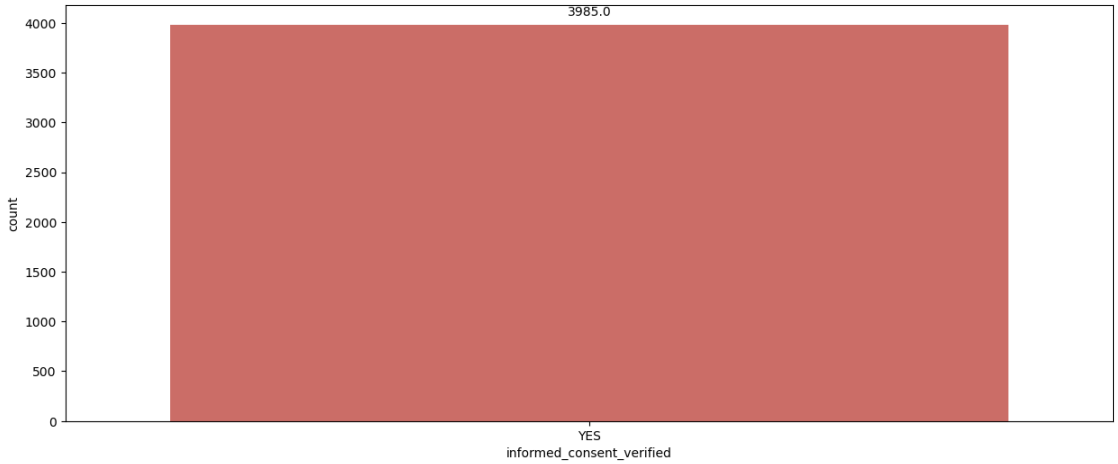
```
has_follow_ups_information
has_follow_ups_information
YES    3046
NO      939
Name: count, dtype: int64

has_drugs_information
has_drugs_information
NO     3087
YES     898
Name: count, dtype: int64

has_radiations_information
has_radiations_information
NO     2966
YES    1019
Name: count, dtype: int64

project
project
TCGA-ESCA    3985
Name: count, dtype: int64

stage_event_system_version
stage_event_system_version
7th    2069
6th    1398
5th     518
Name: count, dtype: int64

primary_pathology_tumor_tissue_site
primary_pathology_tumor_tissue_site
Esophagus    3985
Name: count, dtype: int64

primary_pathology_esophageal_tumor_cental_location
primary_pathology_esophageal_tumor_cental_location
Distal           2807
Mid              1038
Proximal          120
Not Available      20
Name: count, dtype: int64

primary_pathology_esophageal_tumor_involvement_sites
primary_pathology_esophageal_tumor_involvement_sites
Distal           2747
Mid               898
```

```
MidDistal         160
Proximal          120
ProximalMid        40
Not Available      20
Name: count, dtype: int64

primary_pathology_histological_type
primary_pathology_histological_type
Esophagus Adenocarcinoma, NOS        2008
Esophagus Squamous Cell Carcinoma    1977
Name: count, dtype: int64

primary_pathology_columnar_metaplasia_present
primary_pathology_columnar_metaplasia_present
NO               1693
Not Available    1595
YES               697
Name: count, dtype: int64

primary_pathology_neoplasm_histologic_grade
primary_pathology_neoplasm_histologic_grade
G2    1635
G3    1018
GX     933
G1     399
Name: count, dtype: int64

primary_pathology_initial_pathologic_diagnosis_method
primary_pathology_initial_pathologic_diagnosis_method
Endoscopic Biopsy         2368
Other method, specify:     859
Surgical Resection         658
Not Available              100
Name: count, dtype: int64

primary_pathology_lymph_node_metastasis_radiographic_evidence
primary_pathology_lymph_node_metastasis_radiographic_evidence
NO               2109
YES              1039
Not Available     837
Name: count, dtype: int64

primary_pathology_primary_lymph_node_presentation_assessment
primary_pathology_primary_lymph_node_presentation_assessment
YES              3005
NO                660
Not Available     320
Name: count, dtype: int64
```

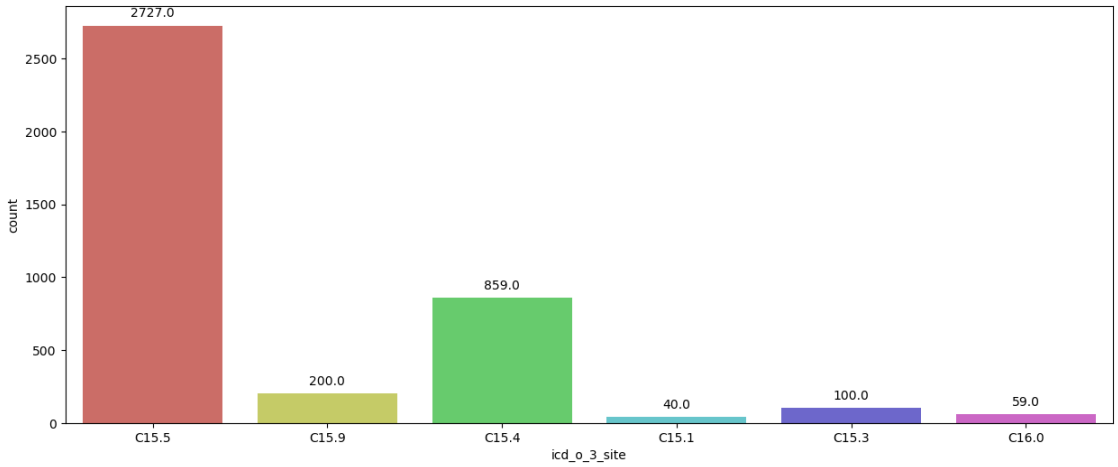
```
primary_pathology_residual_tumor
primary_pathology_residual_tumor
R0                2911
Not Available      520
R1                 298
RX                 216
R2                  40
Name: count, dtype: int64

primary_pathology_radiation_therapy
primary_pathology_radiation_therapy
NO                2949
Not Available      638
YES                398
Name: count, dtype: int64

primary_pathology_postoperative_rx_tx
primary_pathology_postoperative_rx_tx
NO                2988
Not Available      658
YES                339
Name: count, dtype: int64
```

[33]:
```
for i in categorical_features:
    plt.figure(figsize=(15,6))
    ax = sns.countplot(x=i, data=df, palette = 'hls')

    for p in ax.patches:
        height = p.get_height()
        ax.annotate(f'{height}',
                    xy=(p.get_x() + p.get_width() / 2, height),
                    xytext = (0, 10),
                    textcoords = 'offset points',
                    ha = 'center', va = 'center')

    plt.show()
```

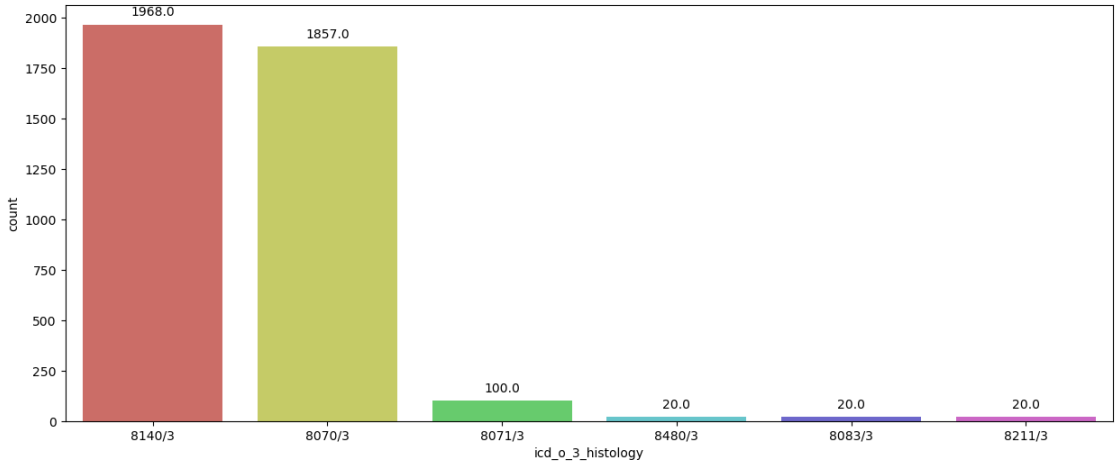
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



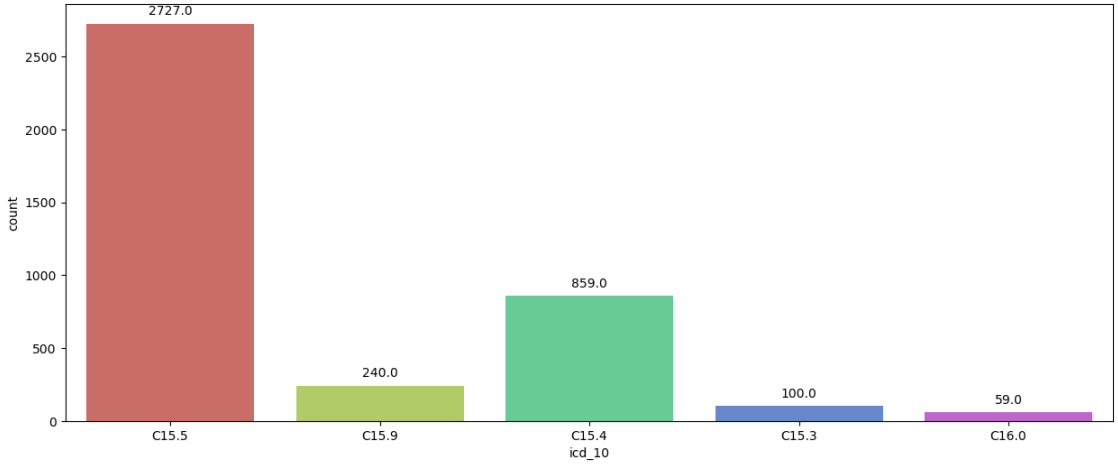

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



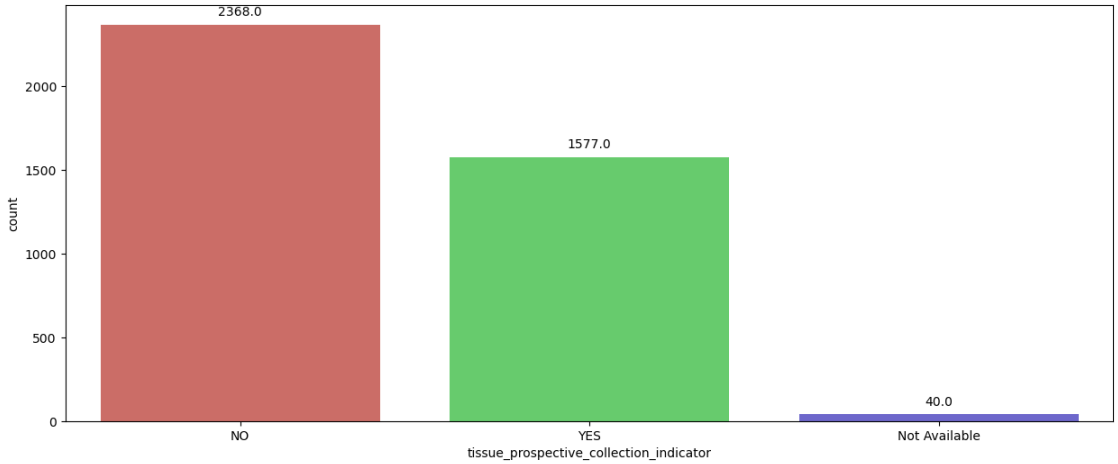

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



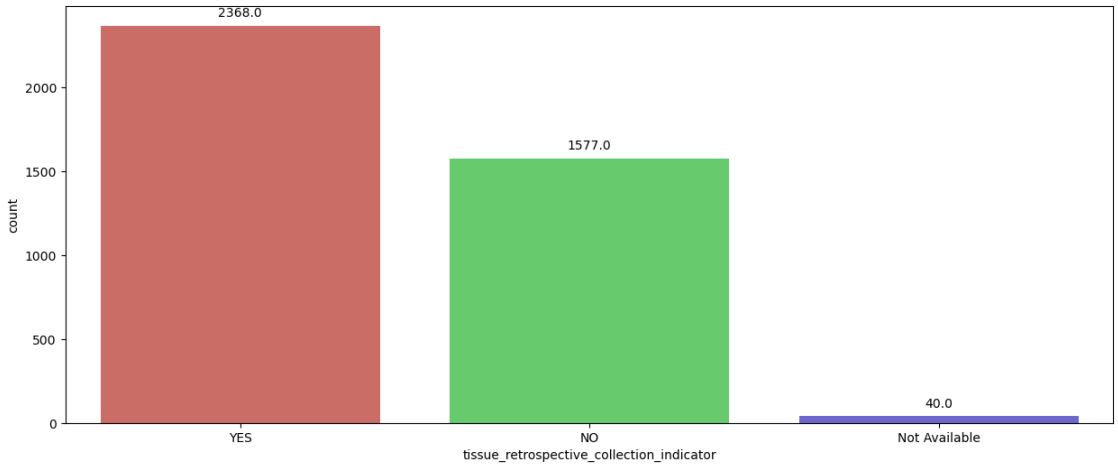

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



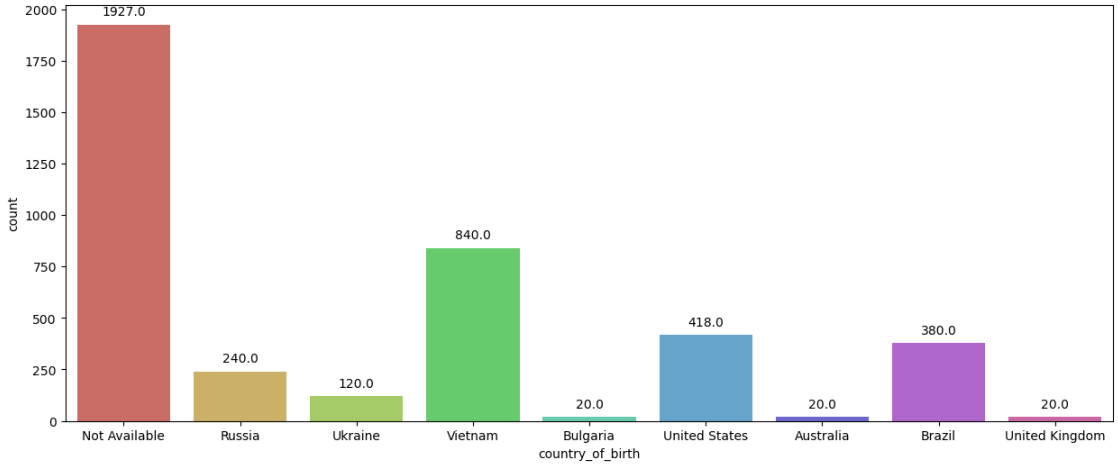

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



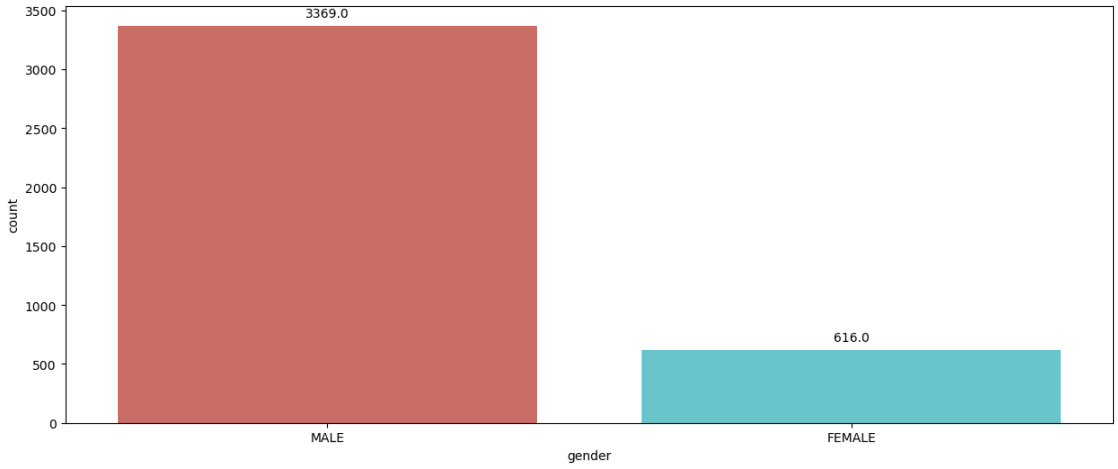

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



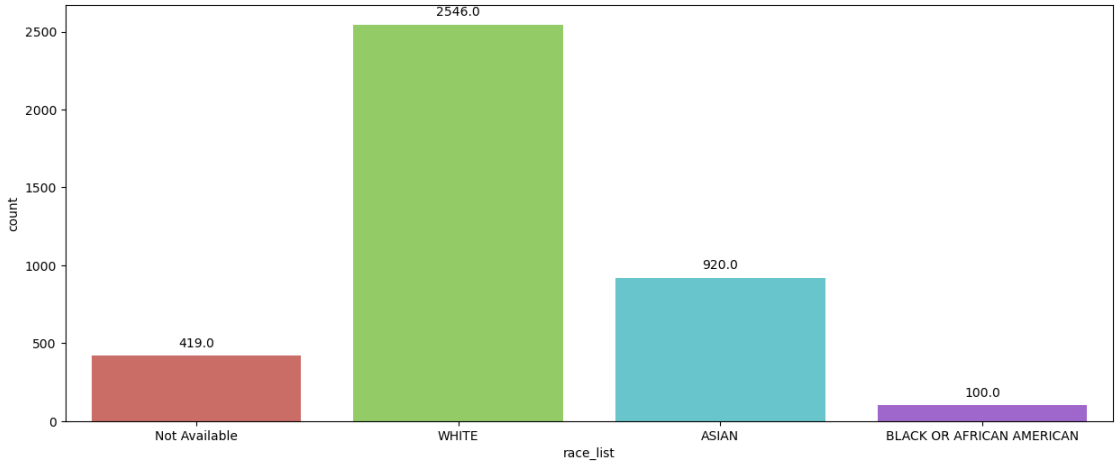

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



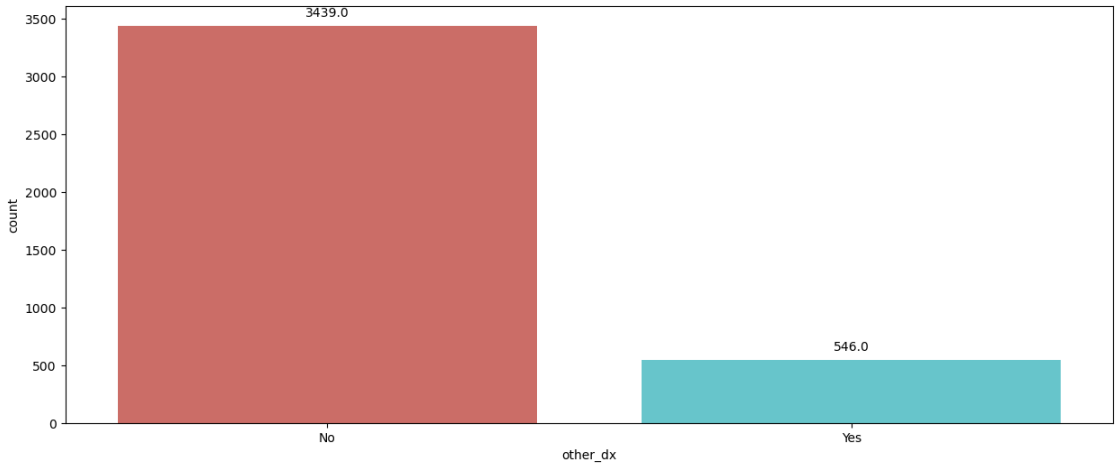

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



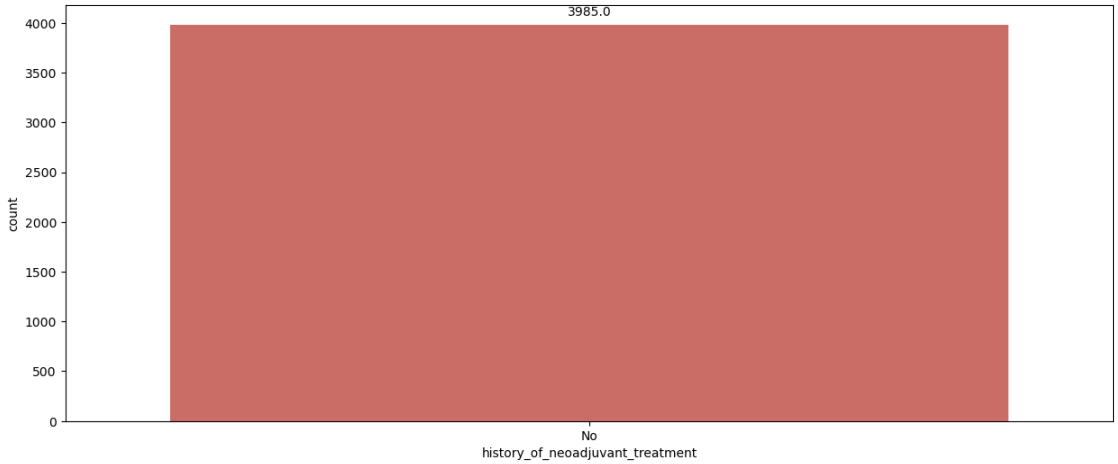

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



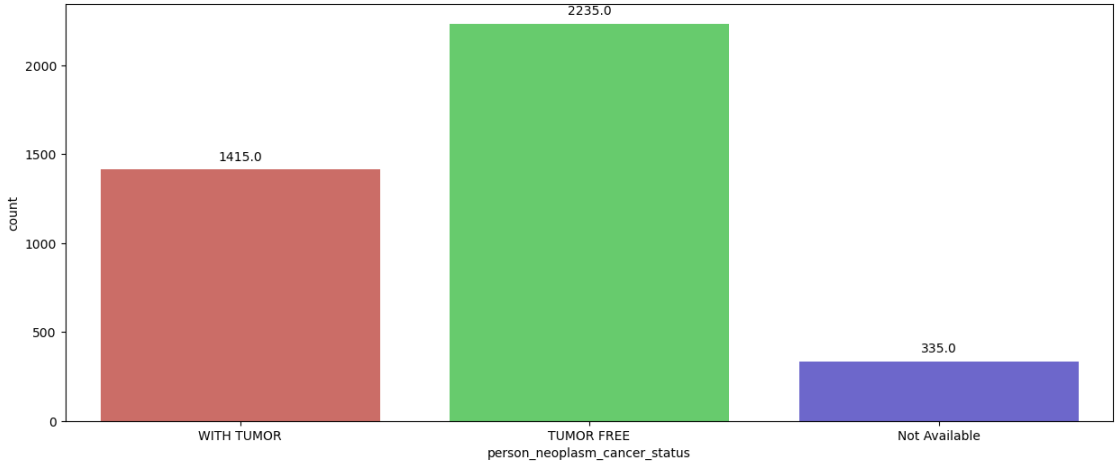

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



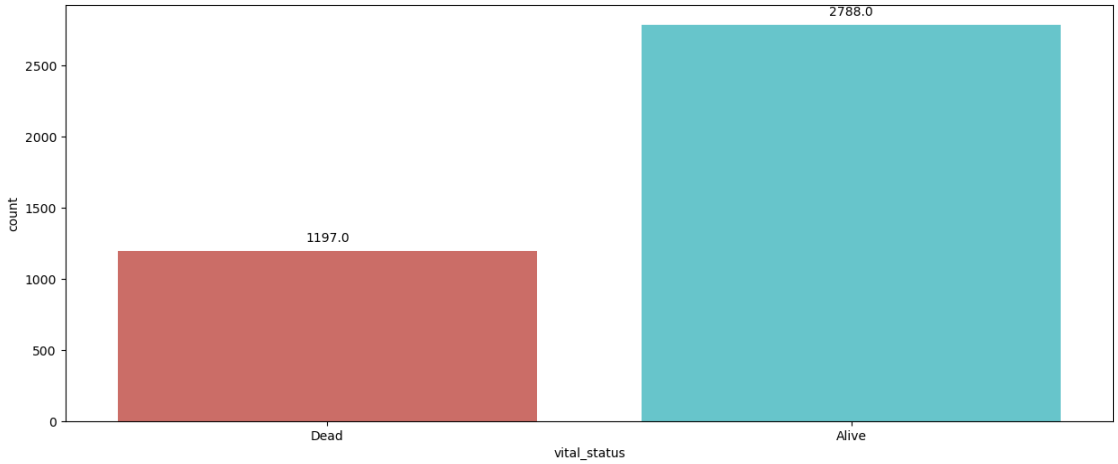

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



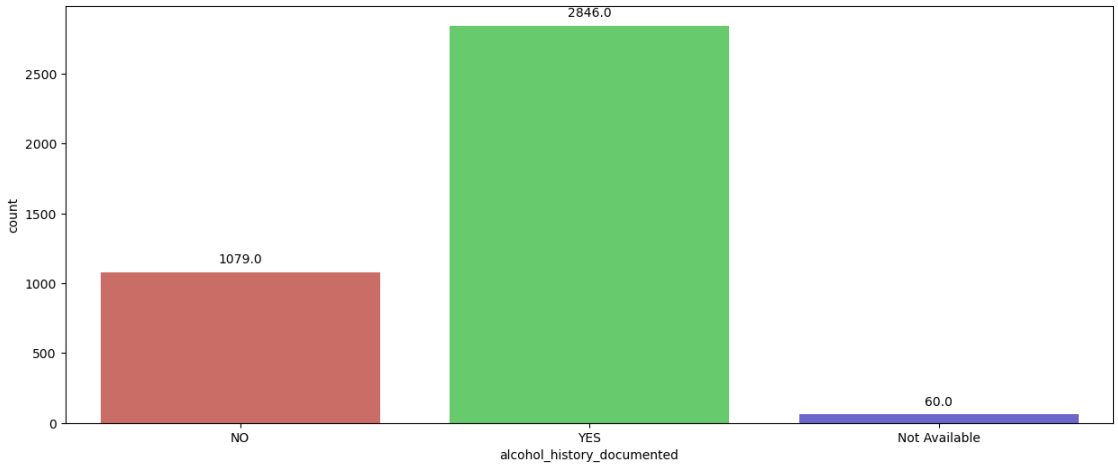

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



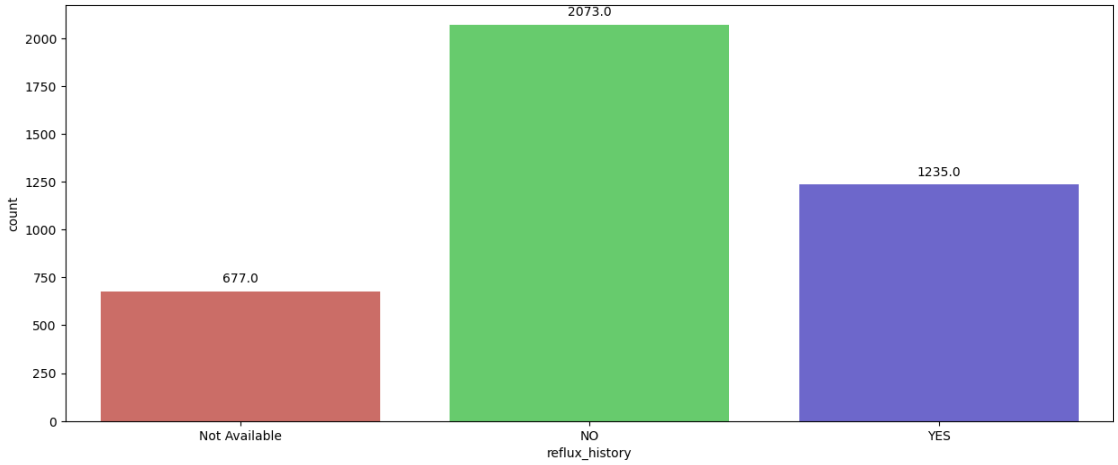

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



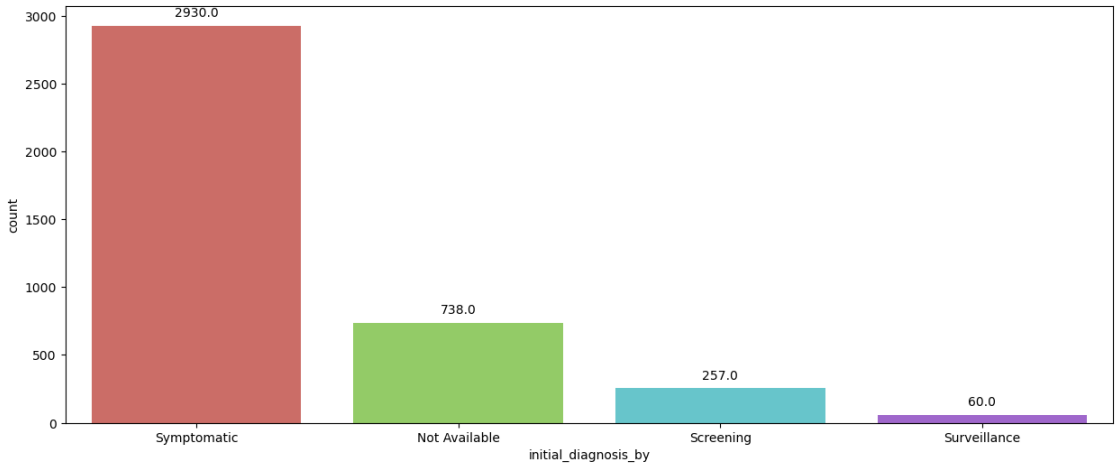

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



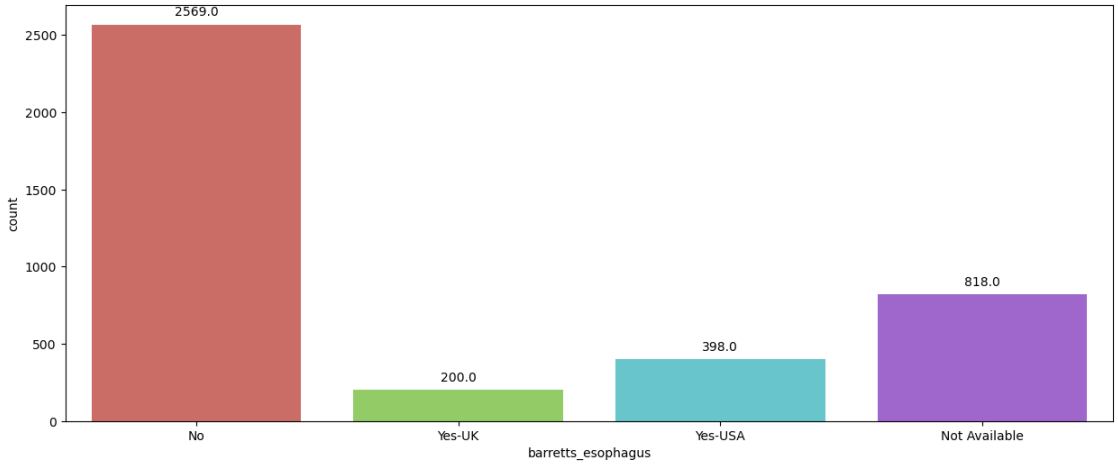

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



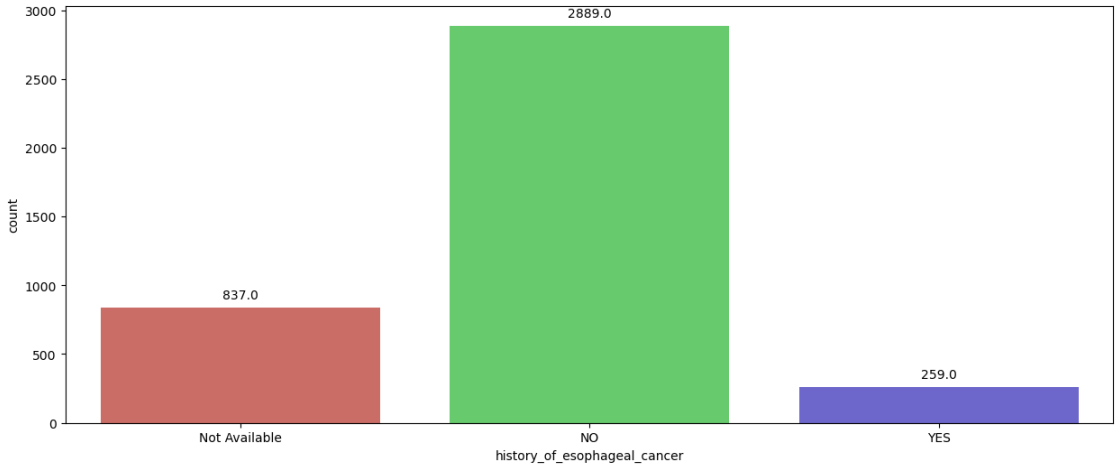

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



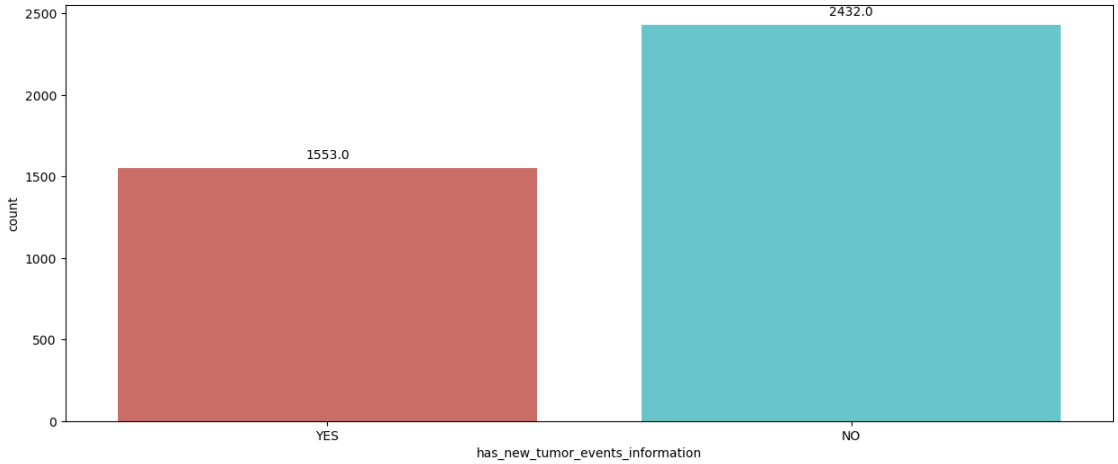

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



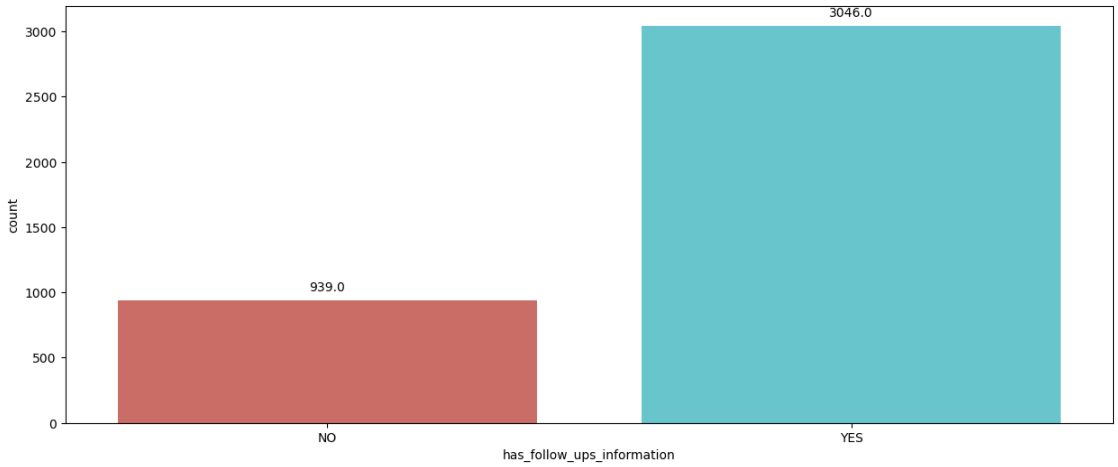

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



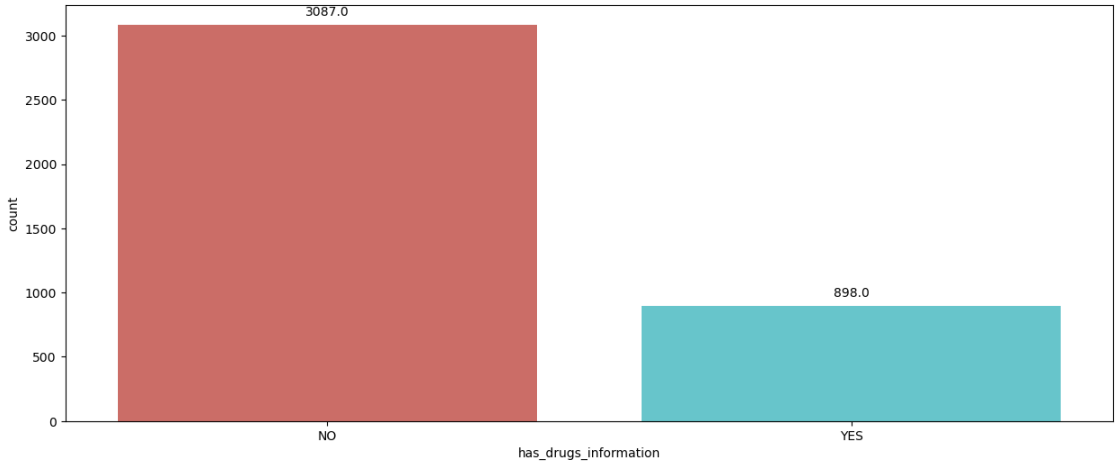

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



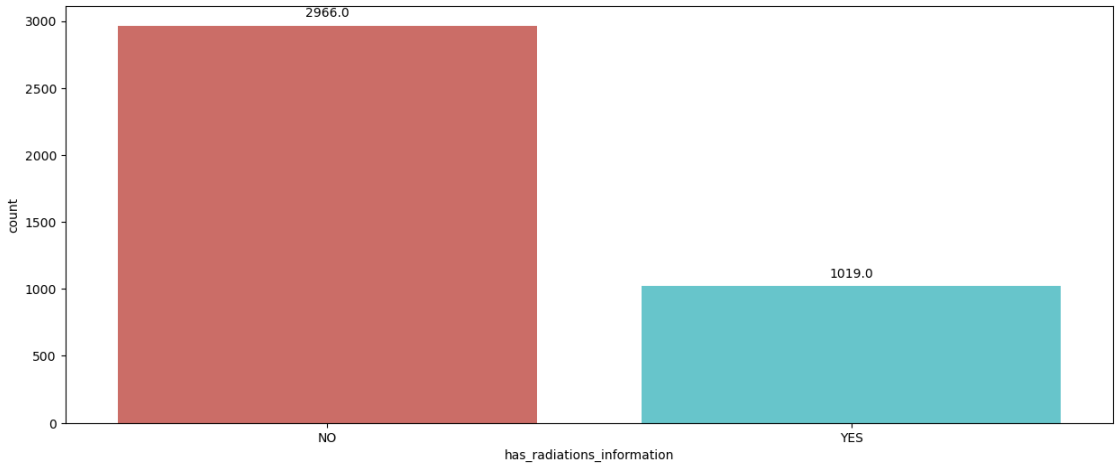

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



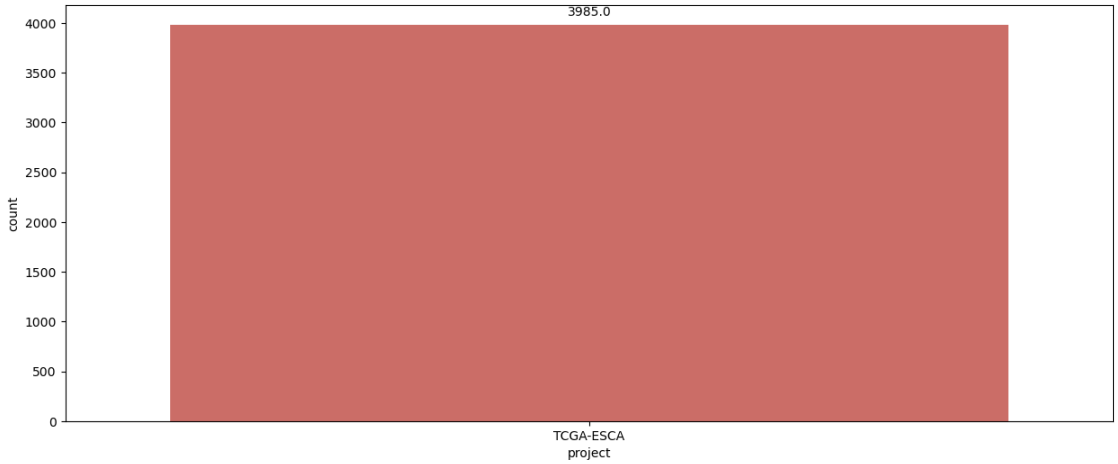

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



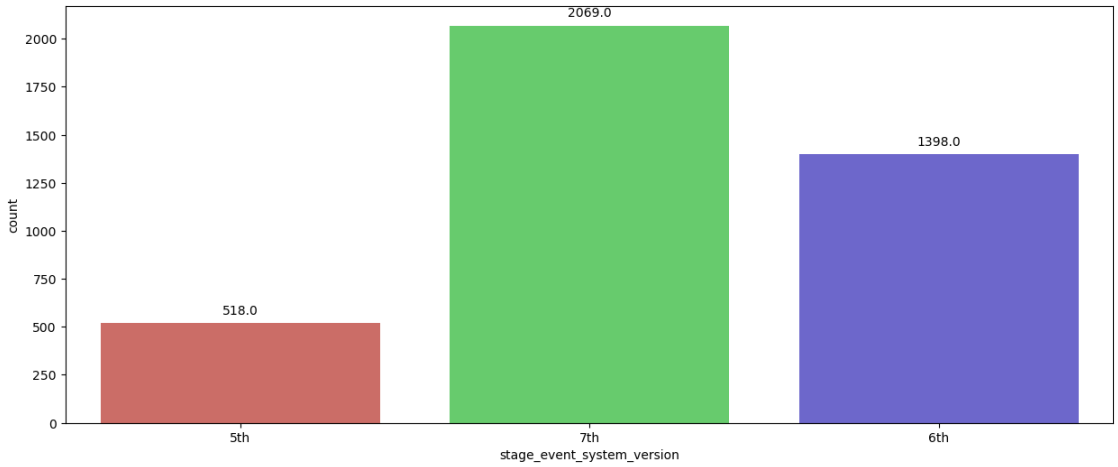

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



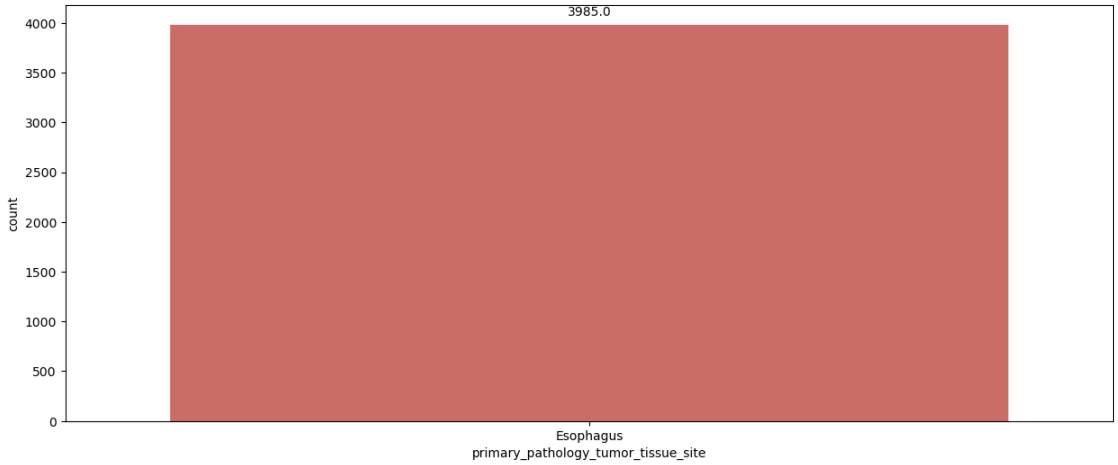

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



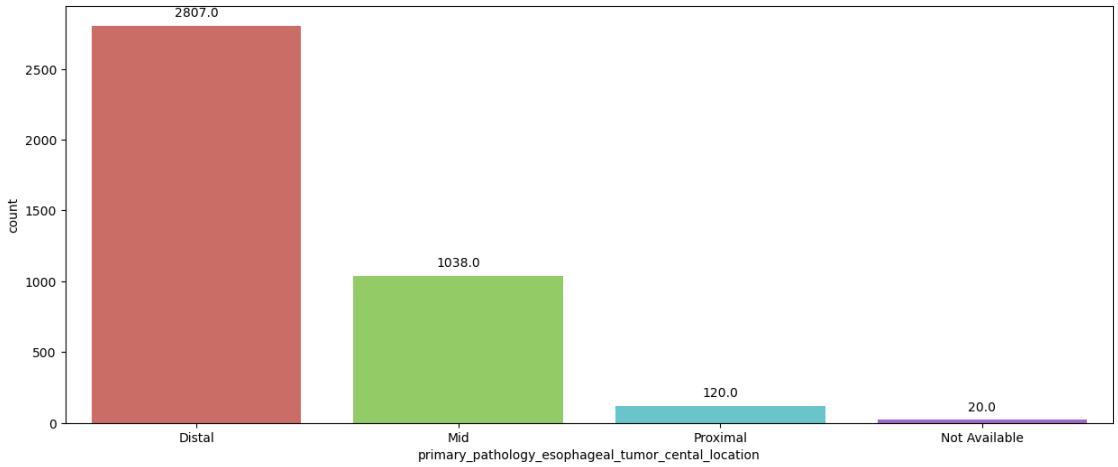

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



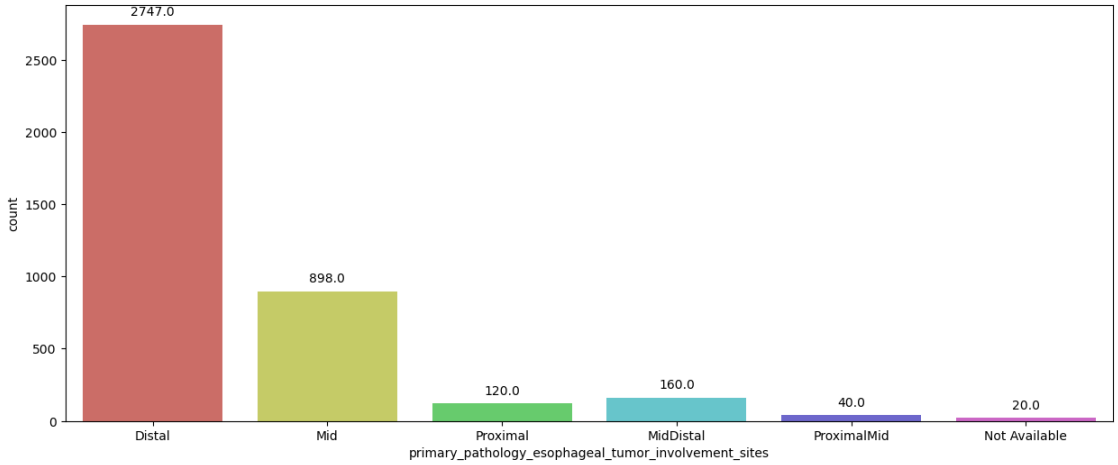

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



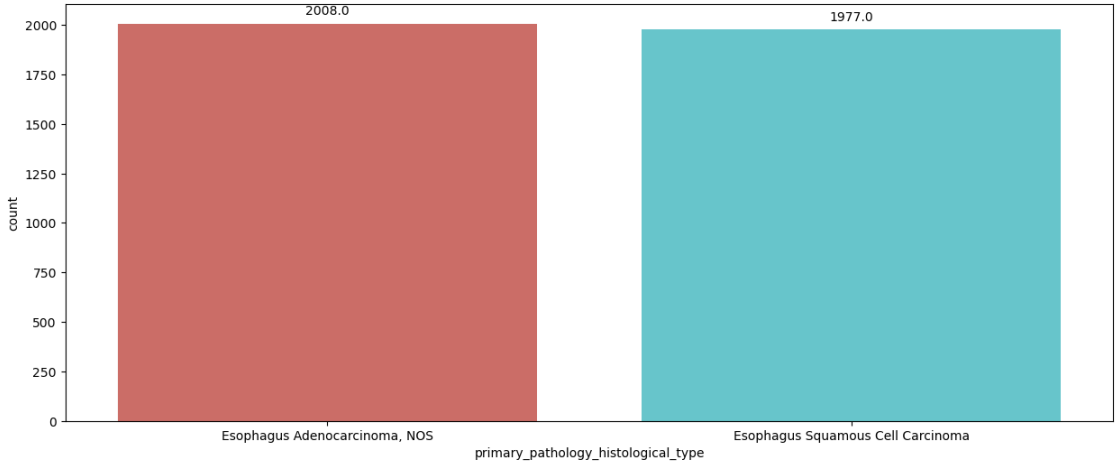

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



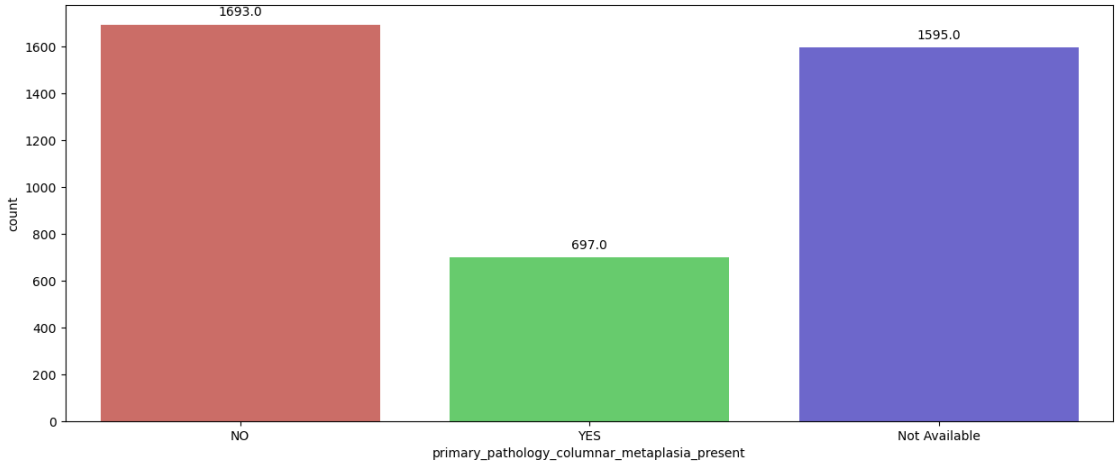

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



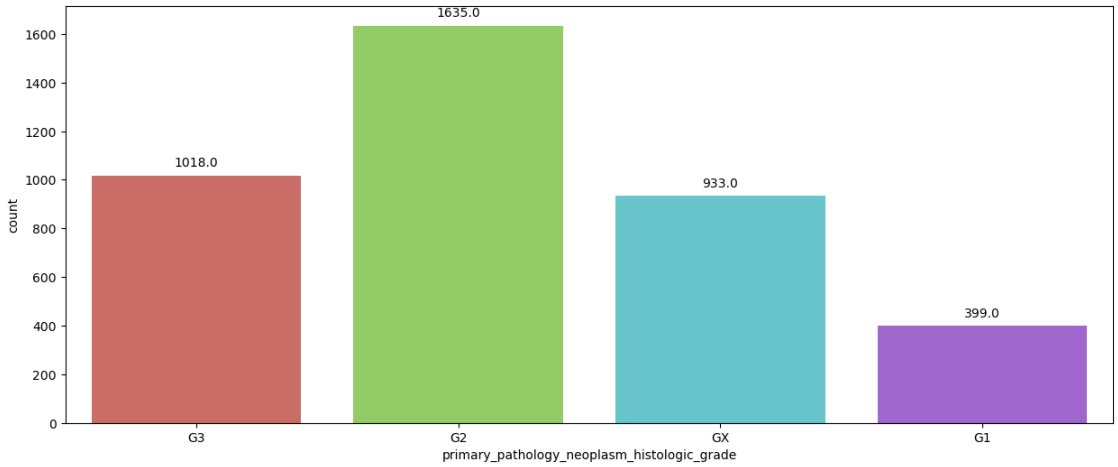

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



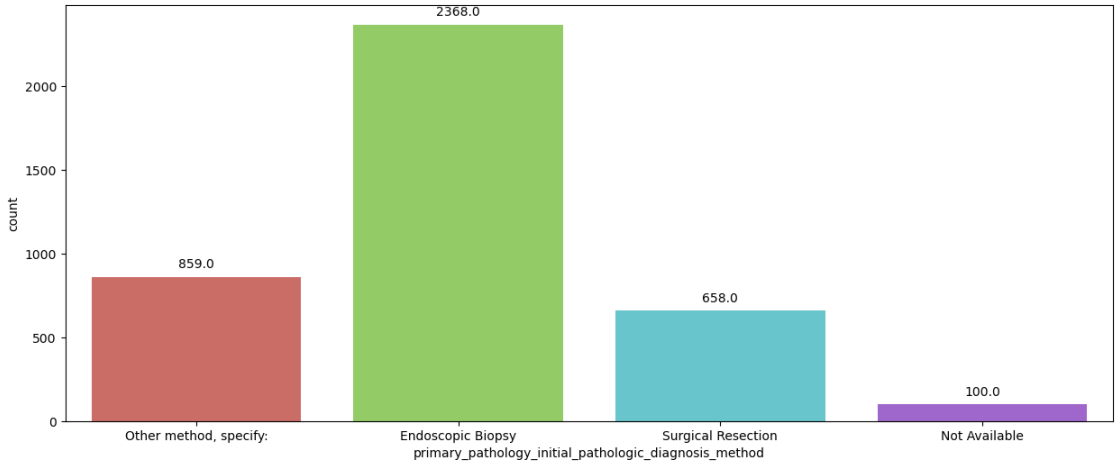

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



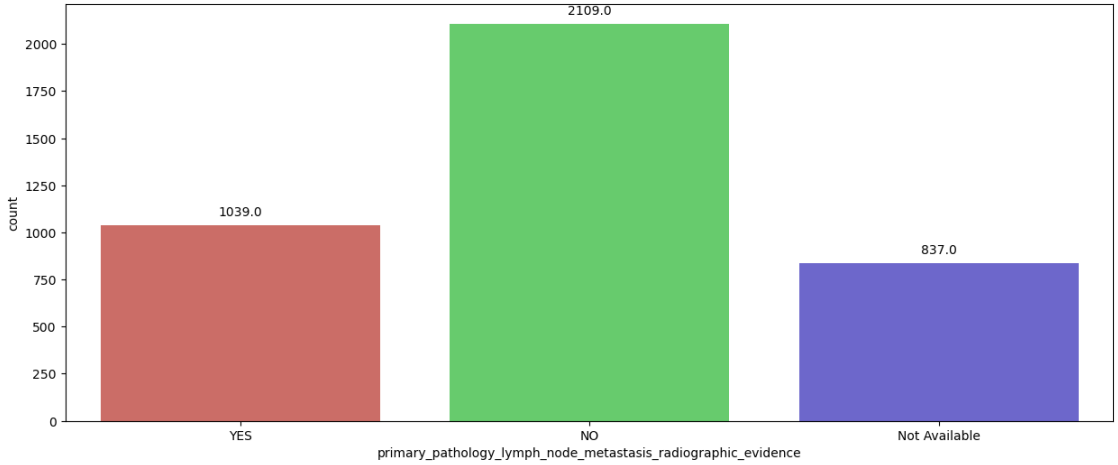

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



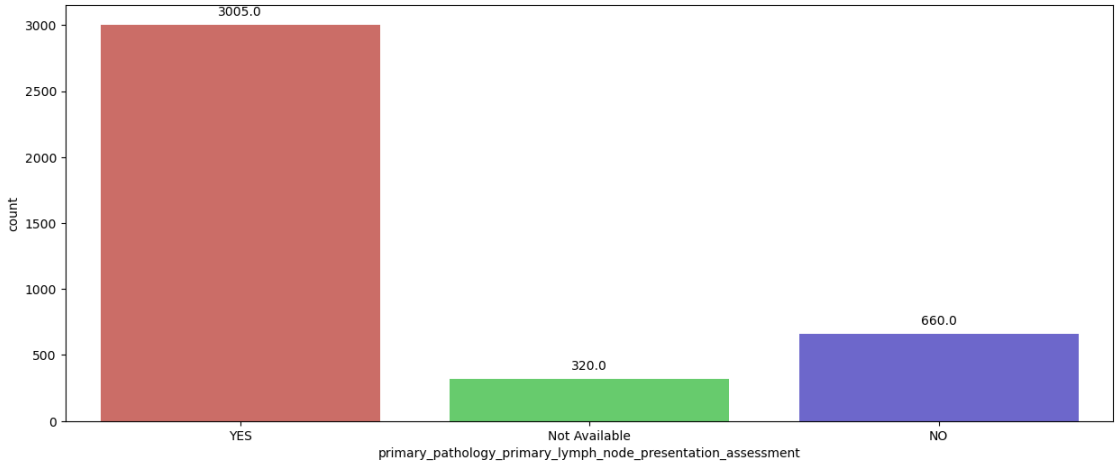

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



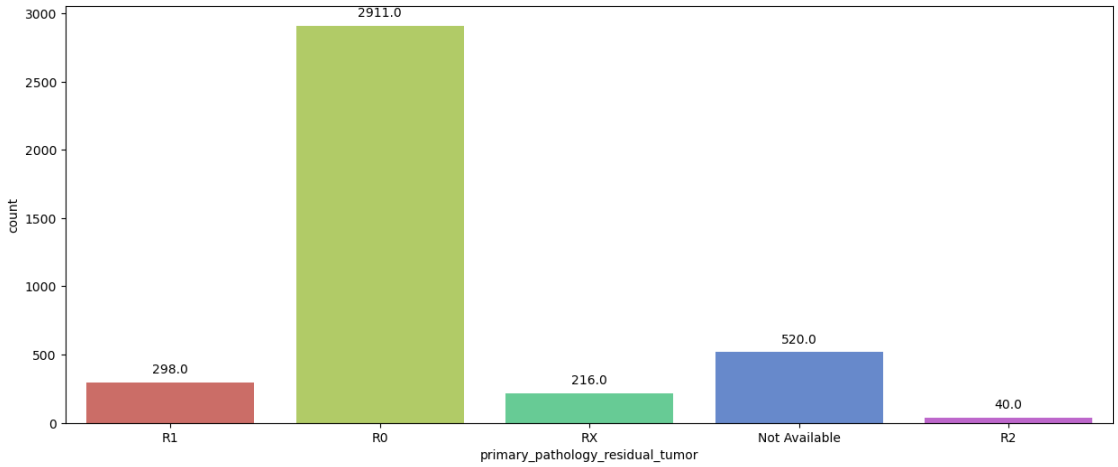

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



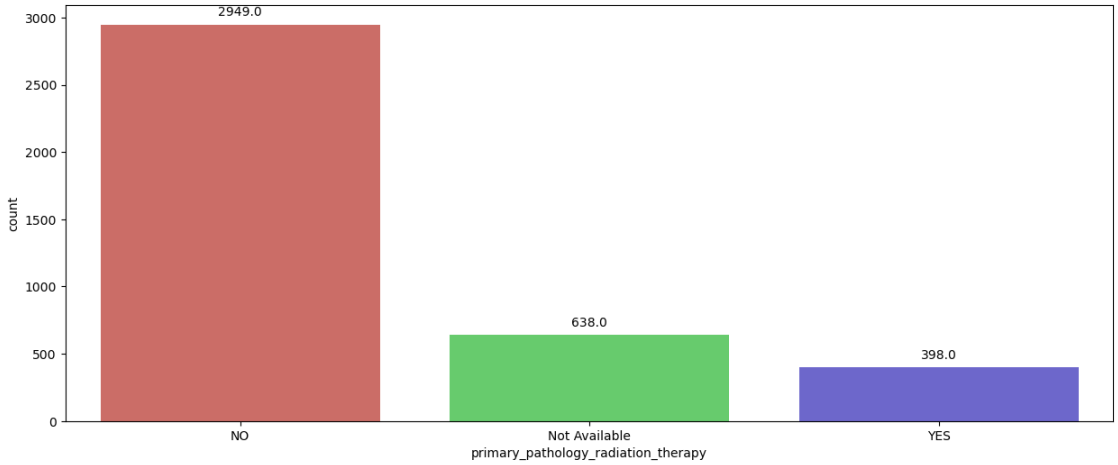

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



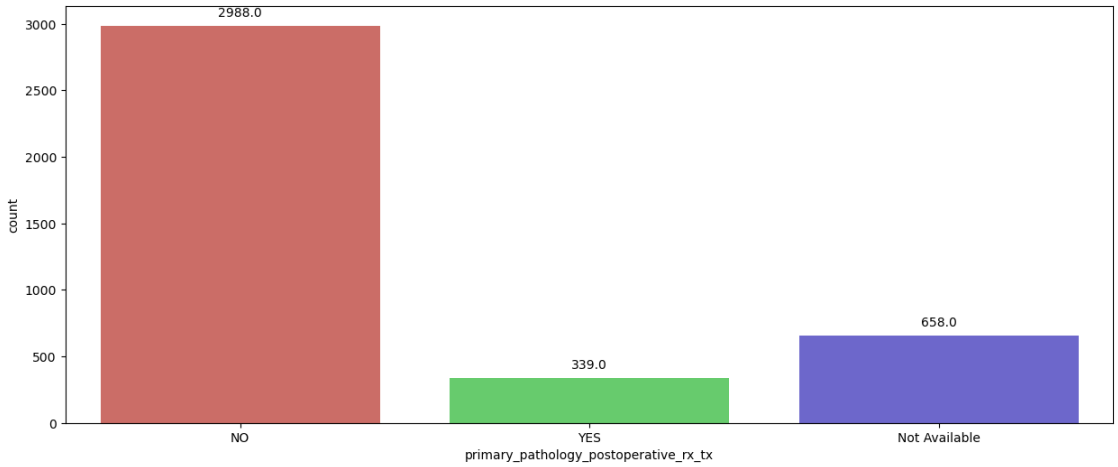

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\531674883.py:3: FutureWarning:

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

  ax = sns.countplot(x=i, data=df, palette = 'hls')
```



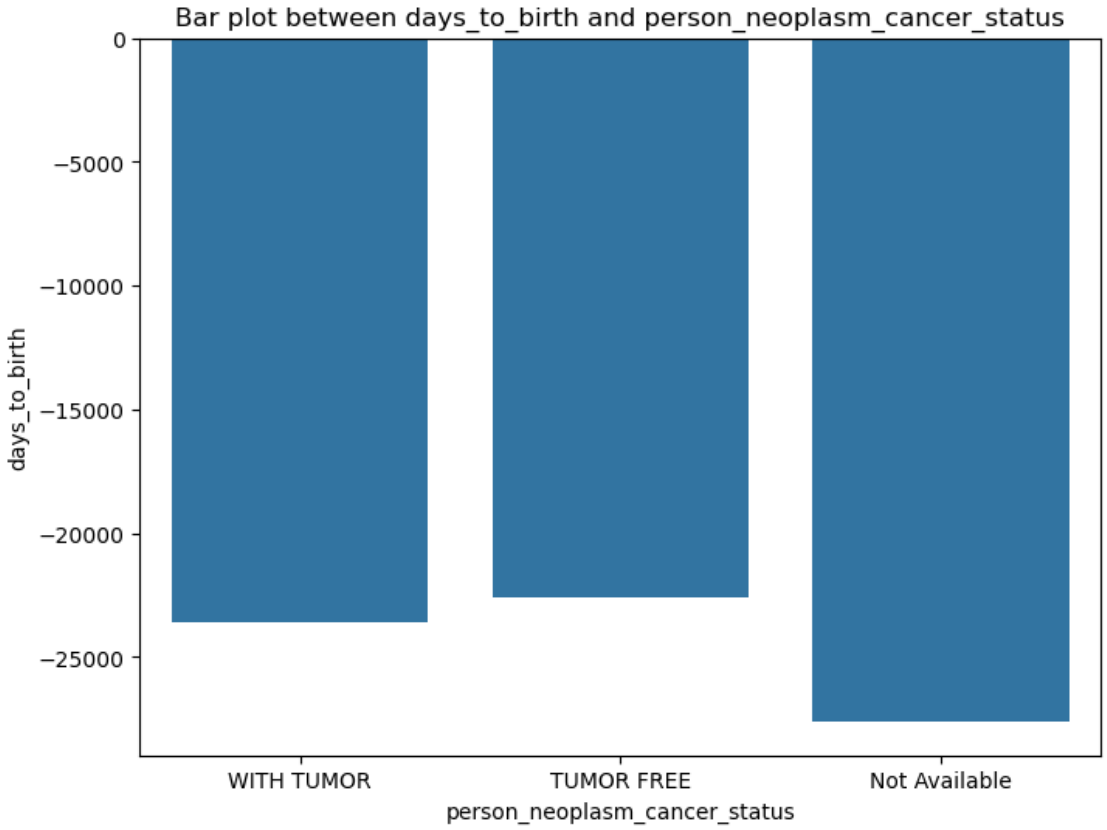
[37]:
```
for i in categorical_features:
    counts = df[i].value_counts()
    fig = px.pie(counts, values=counts.values, names=counts.
 ↪index,title=f'Distribution of {i}')
    fig.show()

from sklearn.feature_selection import chi2
```

```
from sklearn.preprocessing import LabelEncoder

label_enc = LabelEncoder()
df_encoded = pd.DataFrame()
for col in categorical_features:
    df_encoded[col] = label_enc.fit_transform(df[col].astype(str))

chi_scores = chi2(df_encoded, df['person_neoplasm_cancer_status'])[0]
chi_scores_series = pd.Series(chi_scores, index=categorical_features).
 ↪sort_values(ascending = False)
best_categorical_features = chi_scores_series[chi_scores_series > 10].index.
 ↪tolist()

best_categorical_features
```

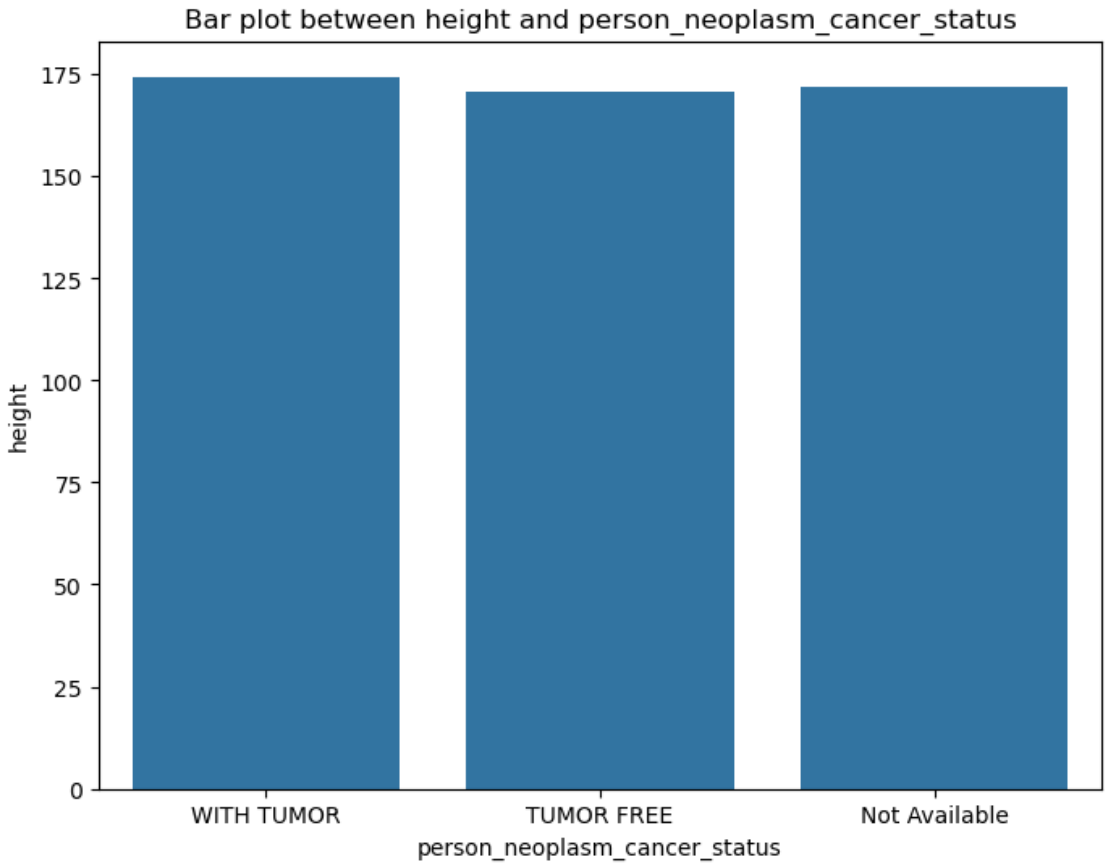
[37]:
```
['person_neoplasm_cancer_status',
 'tissue_prospective_collection_indicator',
 'icd_o_3_histology',
 'vital_status',
 'tissue_retrospective_collection_indicator',
 'country_of_birth',
 'has_new_tumor_events_information',
 'primary_pathology_initial_pathologic_diagnosis_method',
 'initial_diagnosis_by',
 'race_list',
 'other_dx',
 'primary_pathology_histological_type',
 'has_follow_ups_information',
 'primary_pathology_esophageal_tumor_involvement_sites',
 'has_radiations_information',
 'primary_pathology_esophageal_tumor_cental_location',
 'reflux_history',
 'primary_pathology_radiation_therapy',
 'has_drugs_information',
 'primary_pathology_residual_tumor',
 'primary_pathology_primary_lymph_node_presentation_assessment',
 'primary_pathology_postoperative_rx_tx',
 'barretts_esophagus',
 'stage_event_system_version',
 'history_of_esophageal_cancer',
 'alcohol_history_documented',
 'primary_pathology_lymph_node_metastasis_radiographic_evidence',
 'primary_pathology_columnar_metaplasia_present',
 'icd_o_3_site',
 'icd_10',
 'primary_pathology_neoplasm_histologic_grade']
```

[40]:
```
target_variable = 'person_neoplasm_cancer_status'

for features in continuous_features:
    plt.figure(figsize=(8,6))
    sns.barplot(y=df[features], x=df[target_variable], ci=None)
    plt.title(f'Bar plot between {features} and {target_variable}')
    plt.show()
```

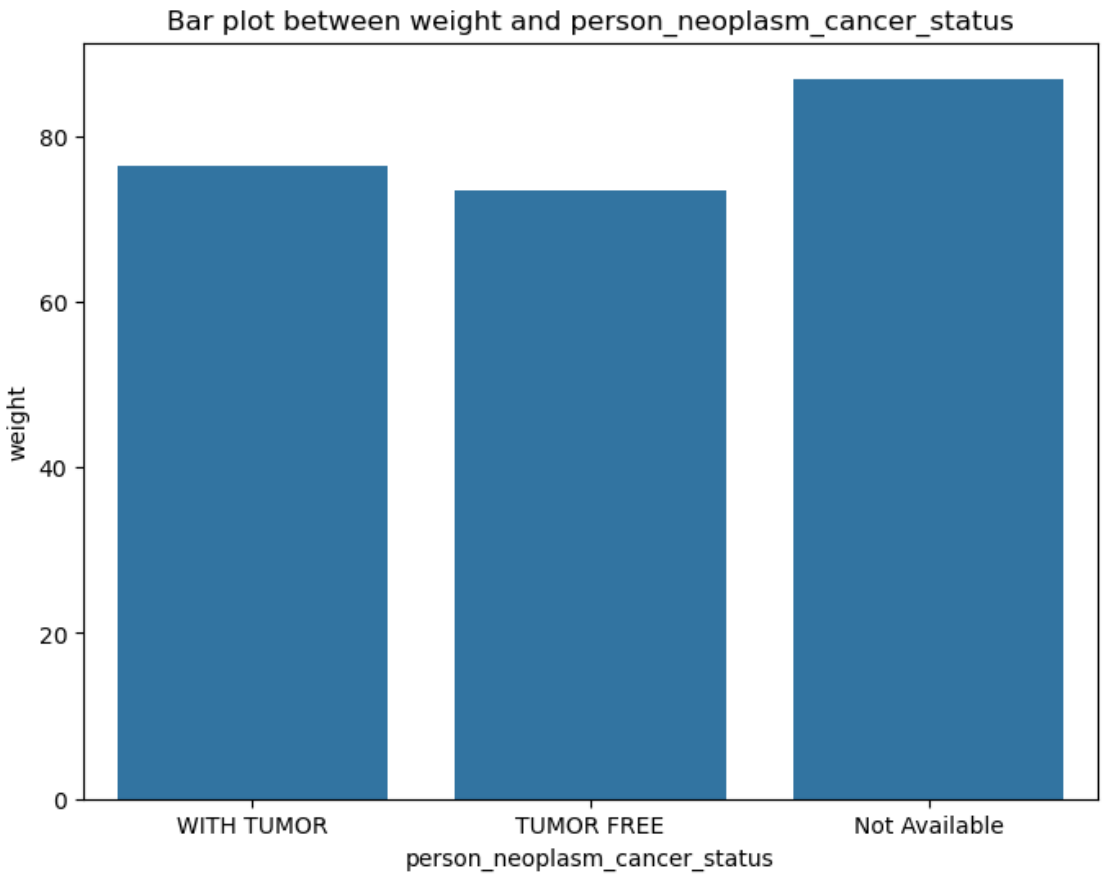
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```



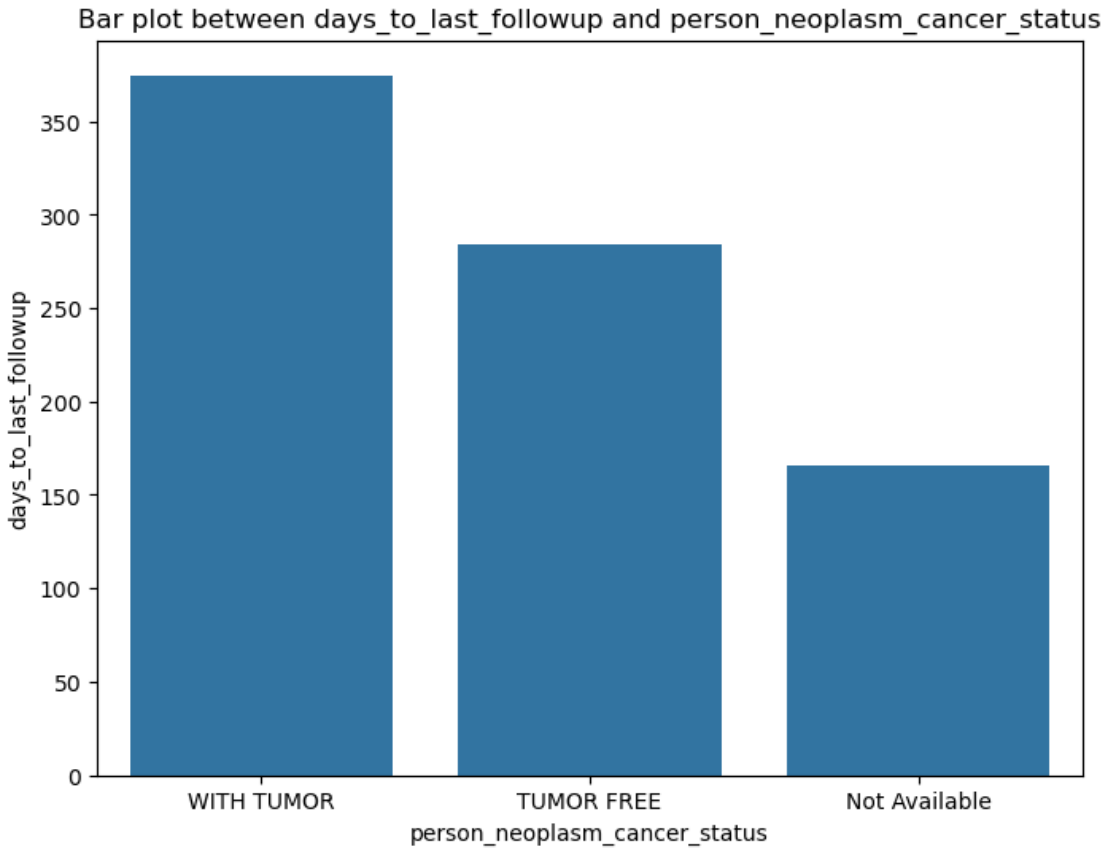

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```



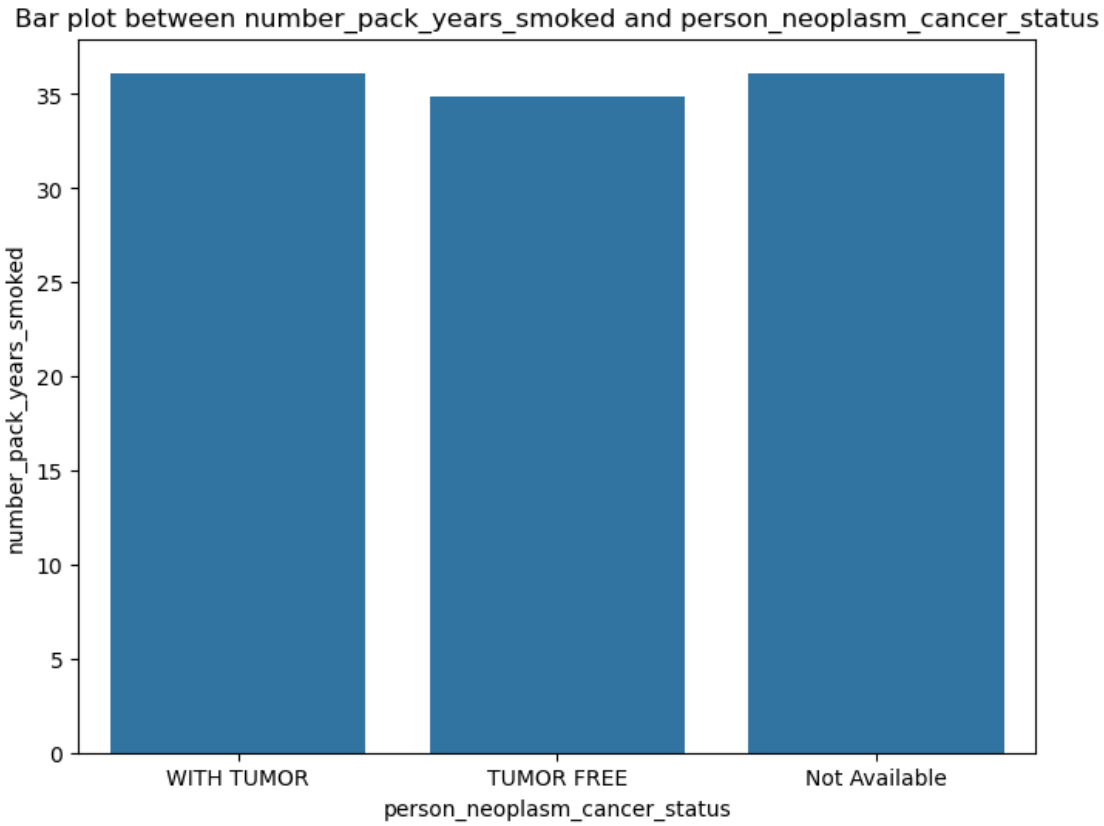

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```



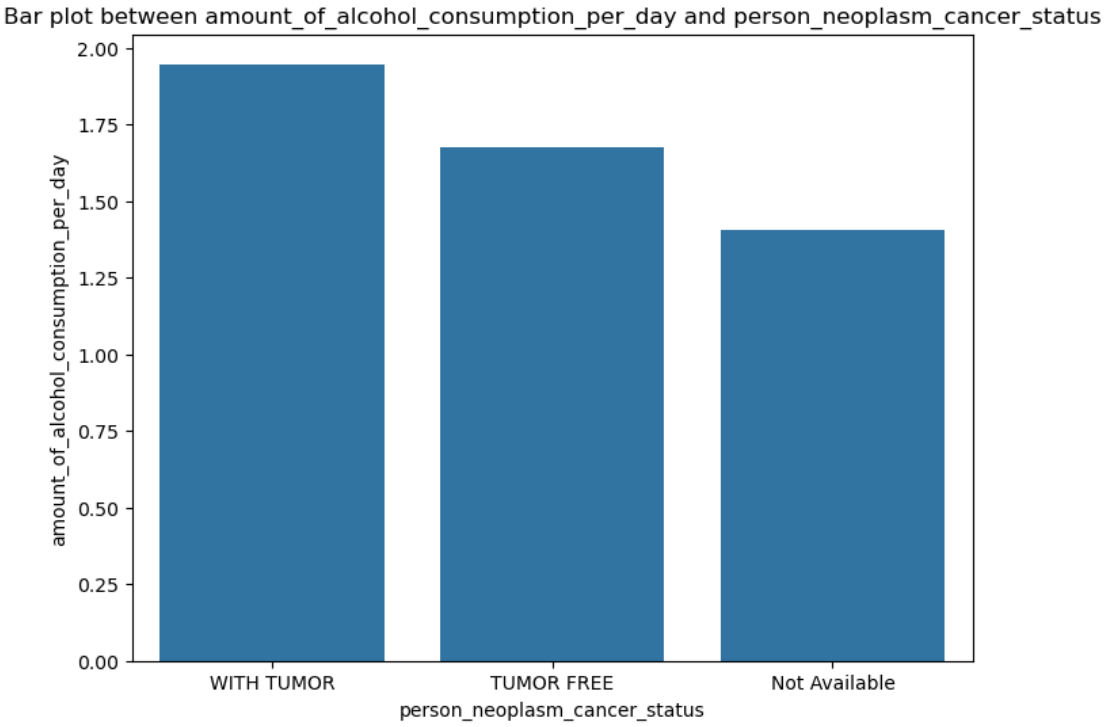

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```



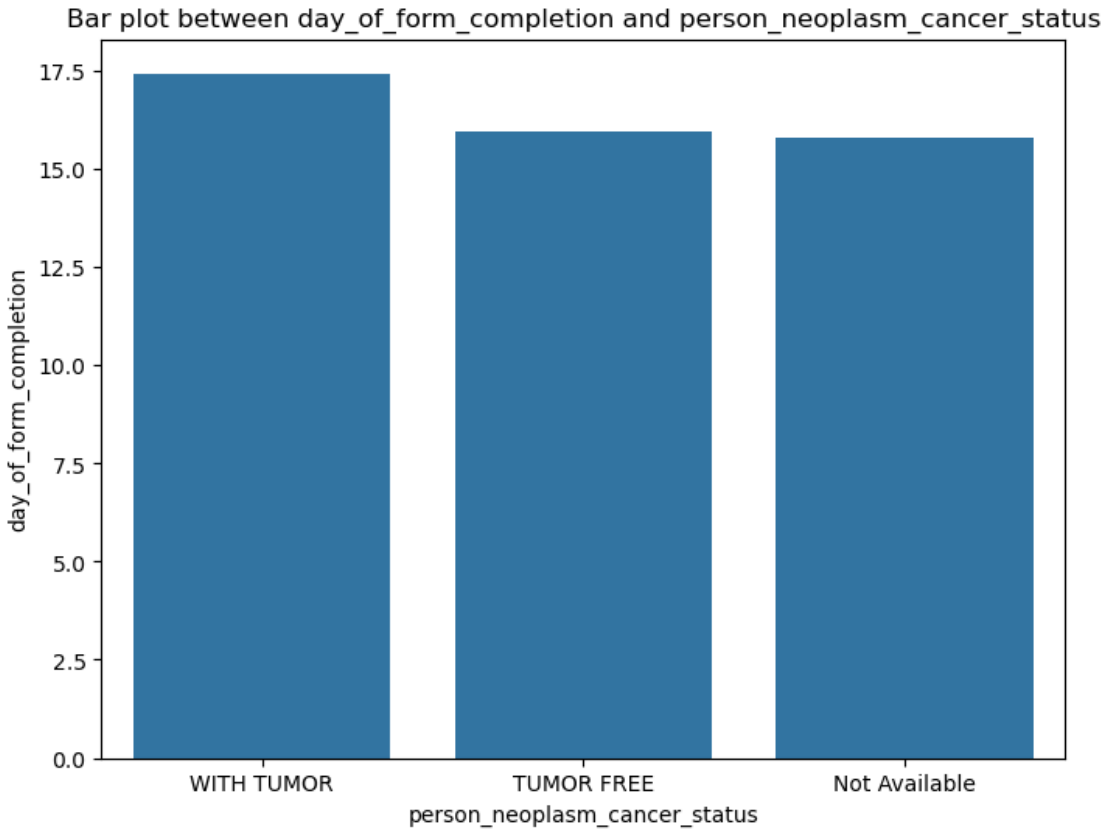

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```



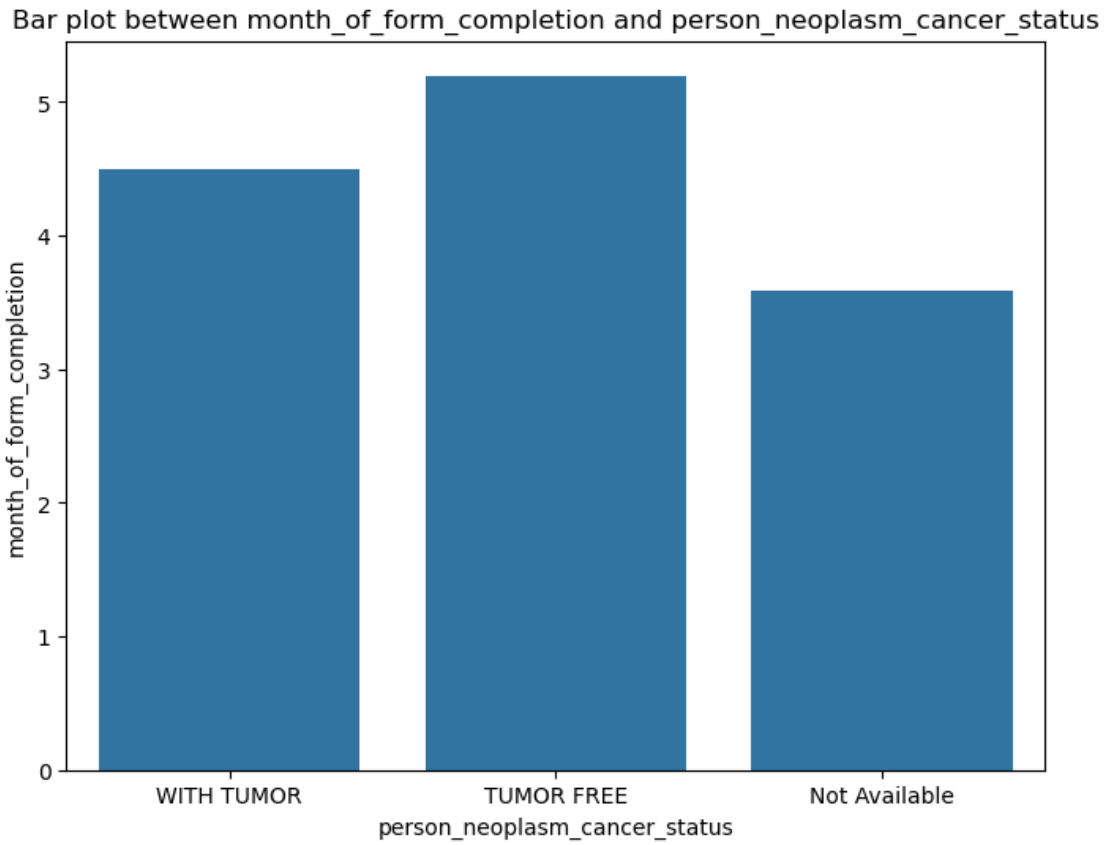

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```



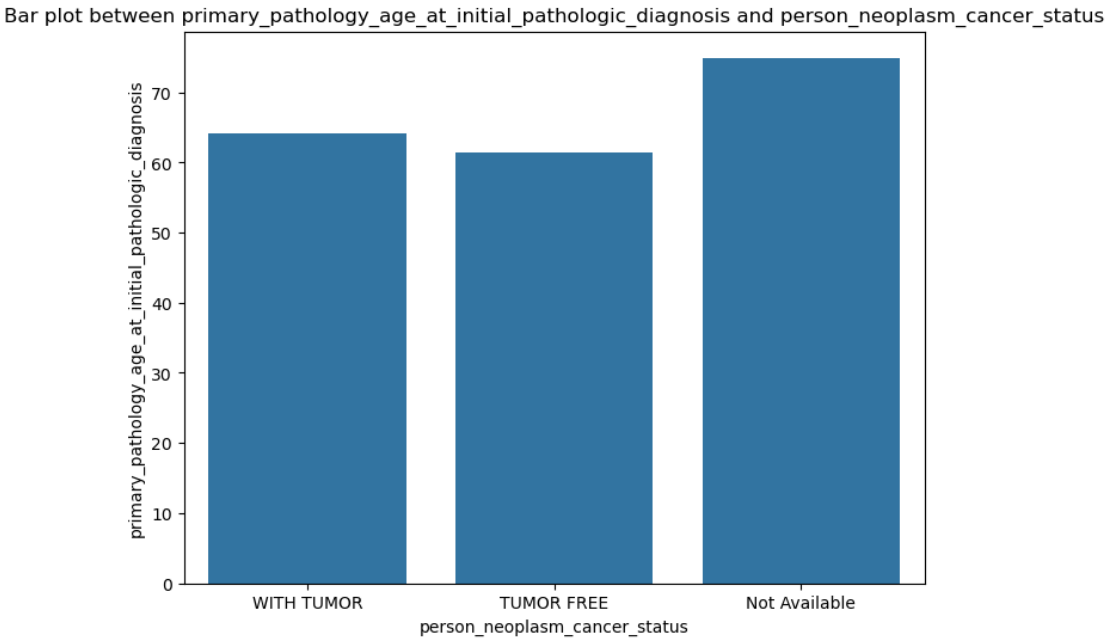

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```




```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```

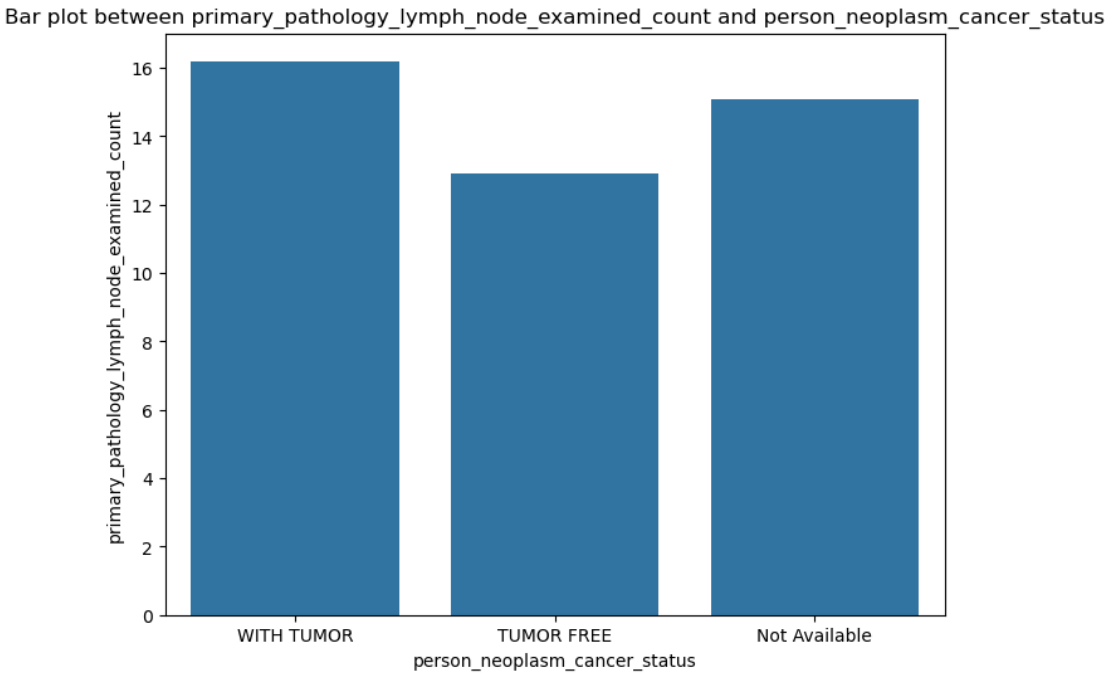



```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```

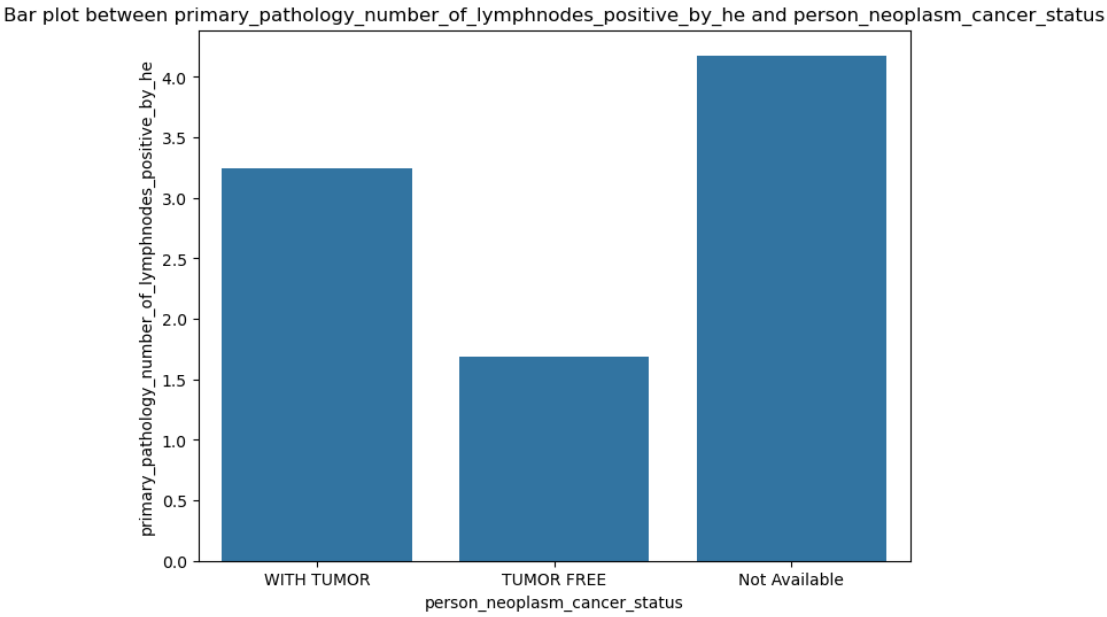



```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```

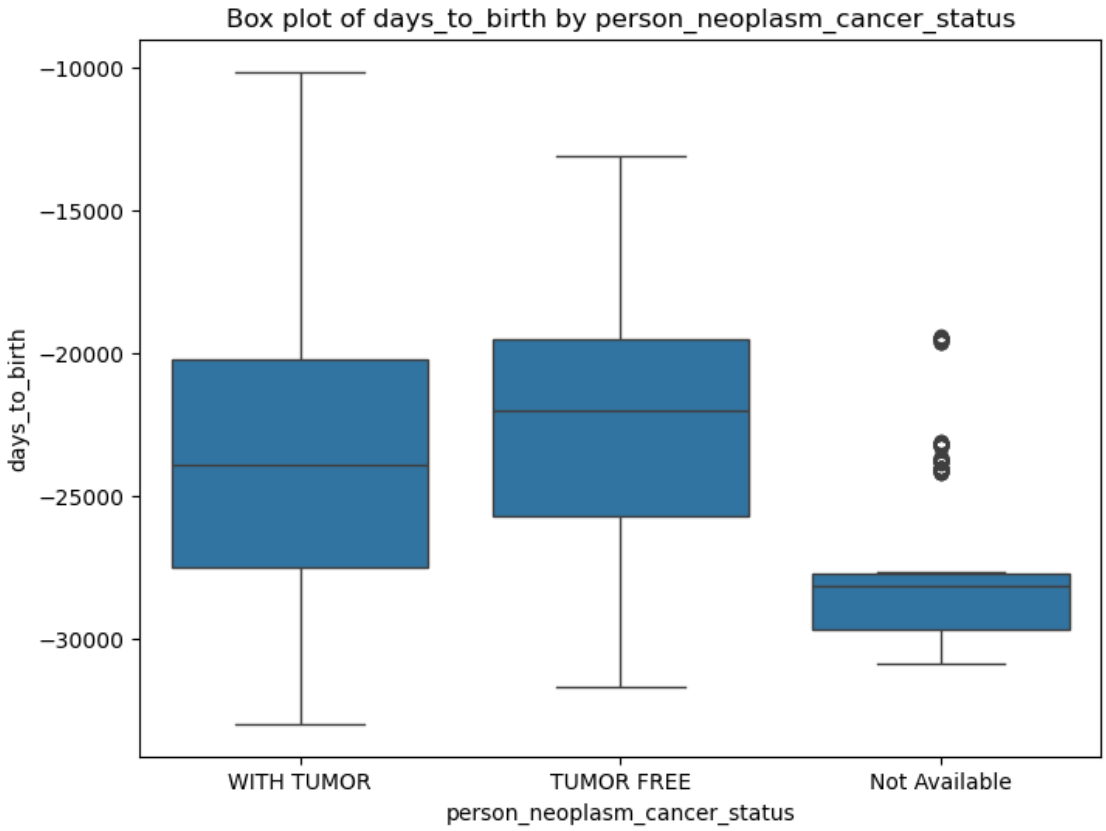
```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```

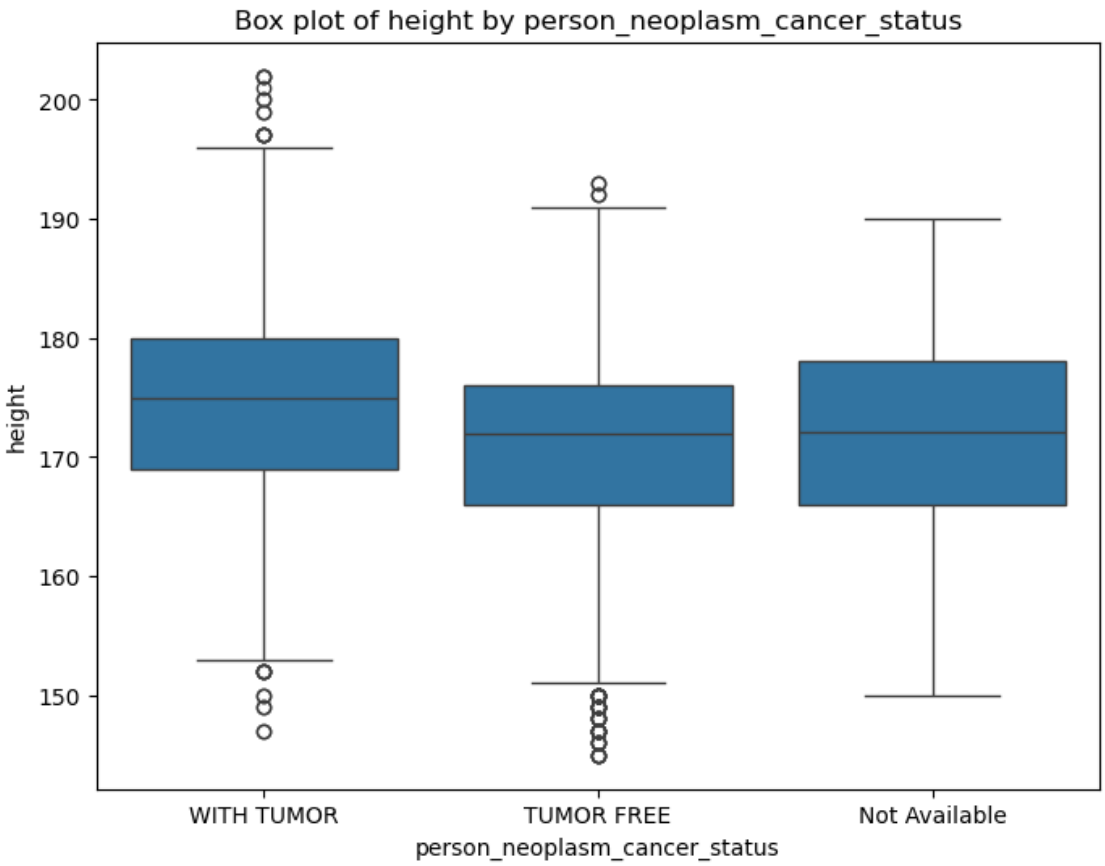


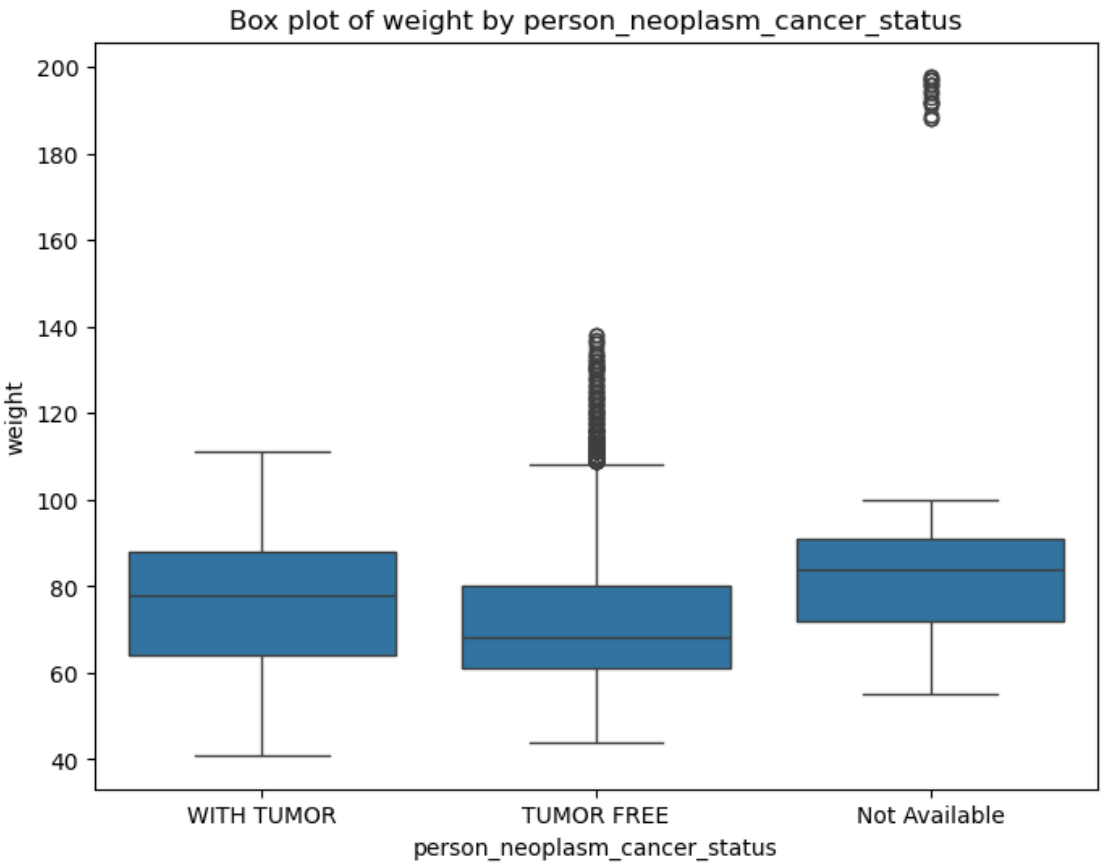

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12948\1311290989.py:5: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.
```

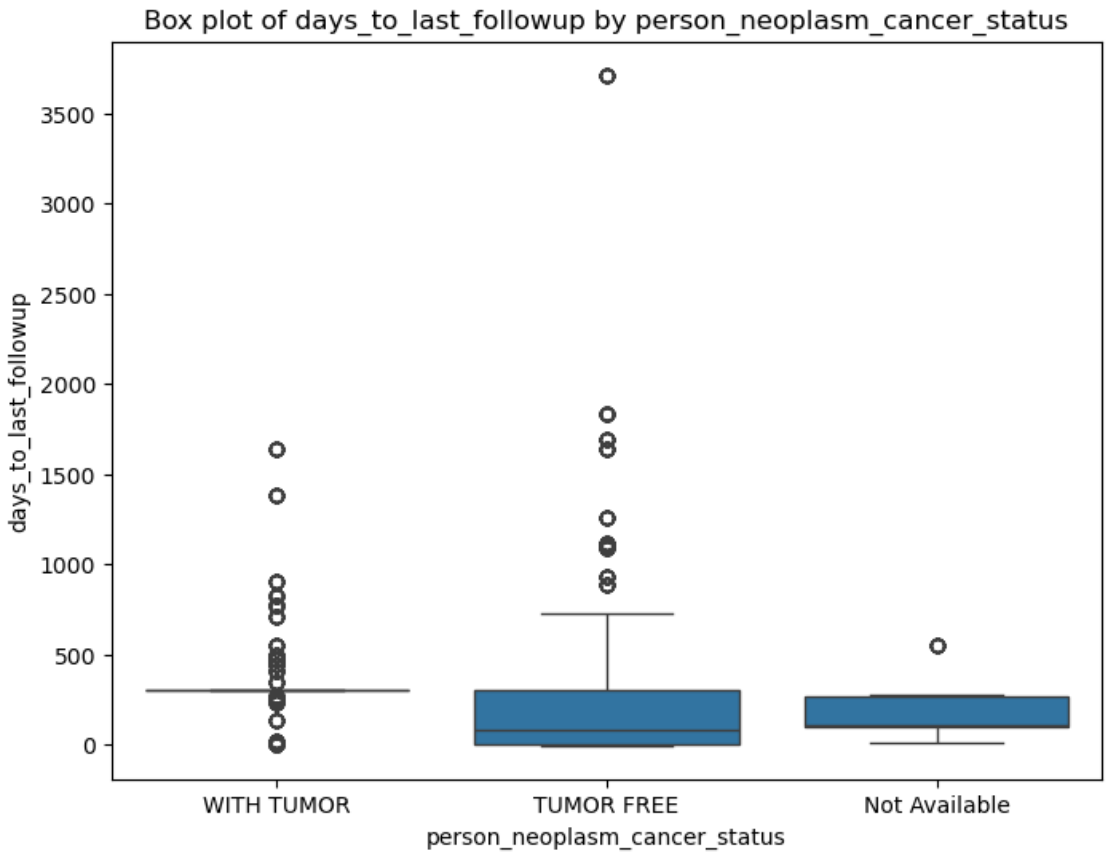


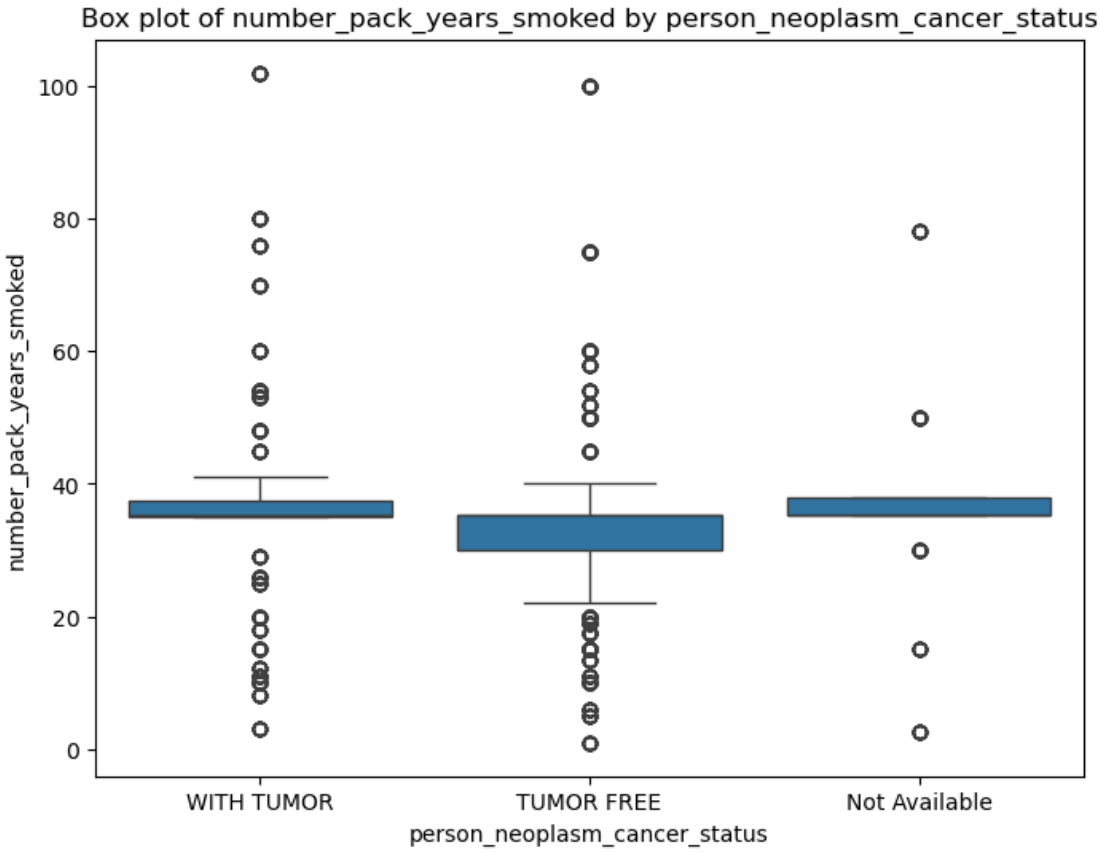
[41]:
```
for features in continuous_features:
    plt.figure(figsize=(8,6))
    sns.boxplot(x=df[target_variable], y=df[features])
    plt.title(f'Box plot of {features} by {target_variable}')
    plt.show()
```

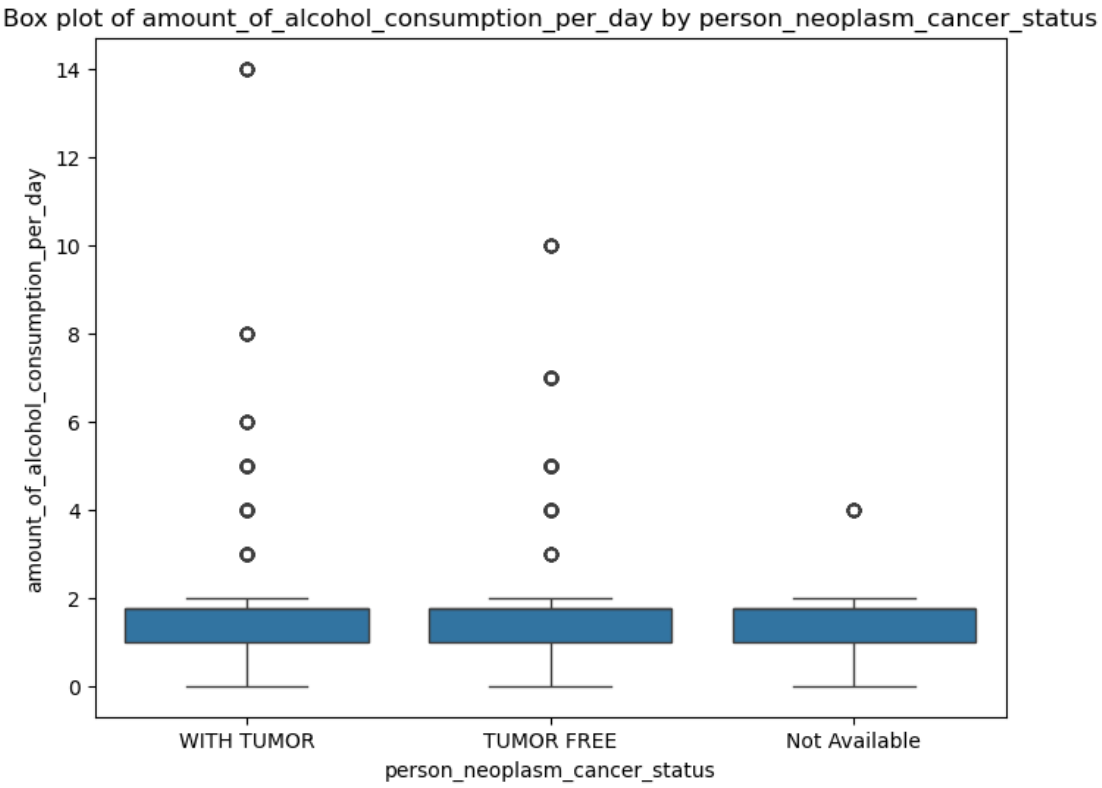

Box plot of days_to_birth by person_neoplasm_cancer_status
−10000
−15000
−20000
−25000
−30000
days_to_birth
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status

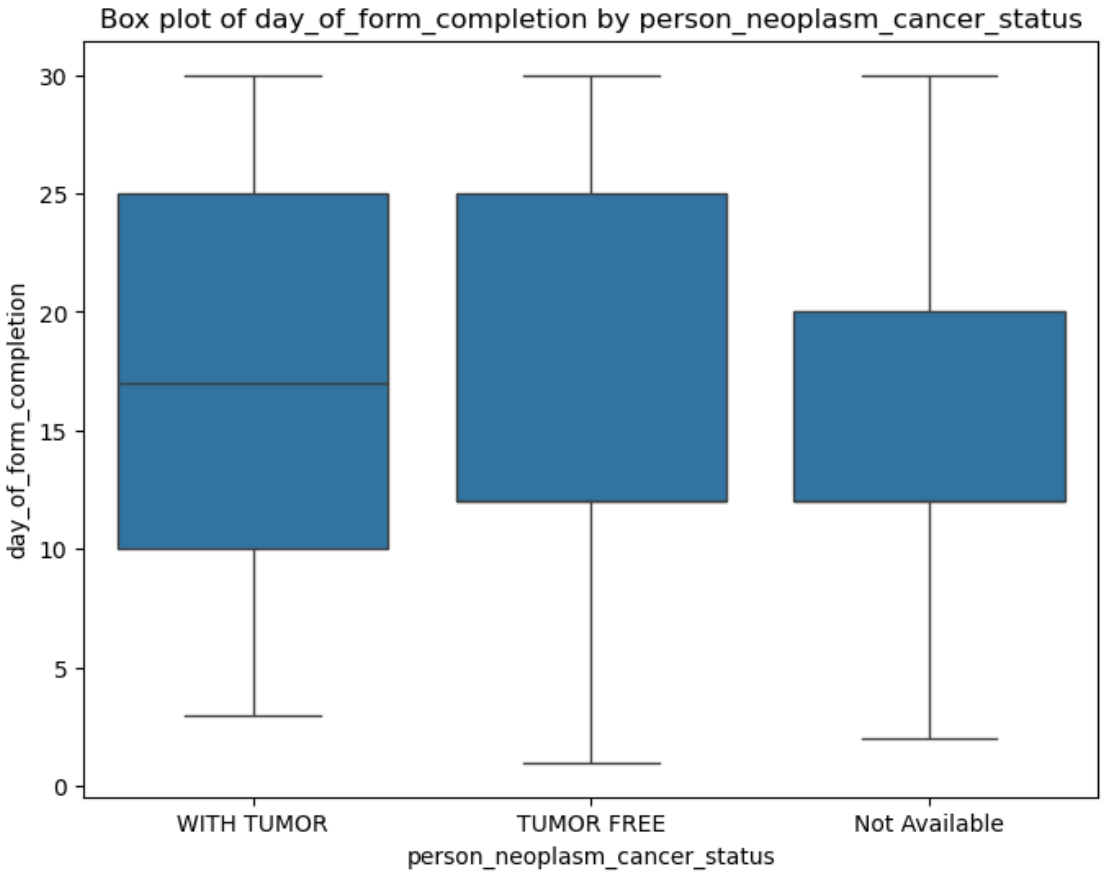

Box plot of height by person_neoplasm_cancer_status
height
200
190
180
170
160
150
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status

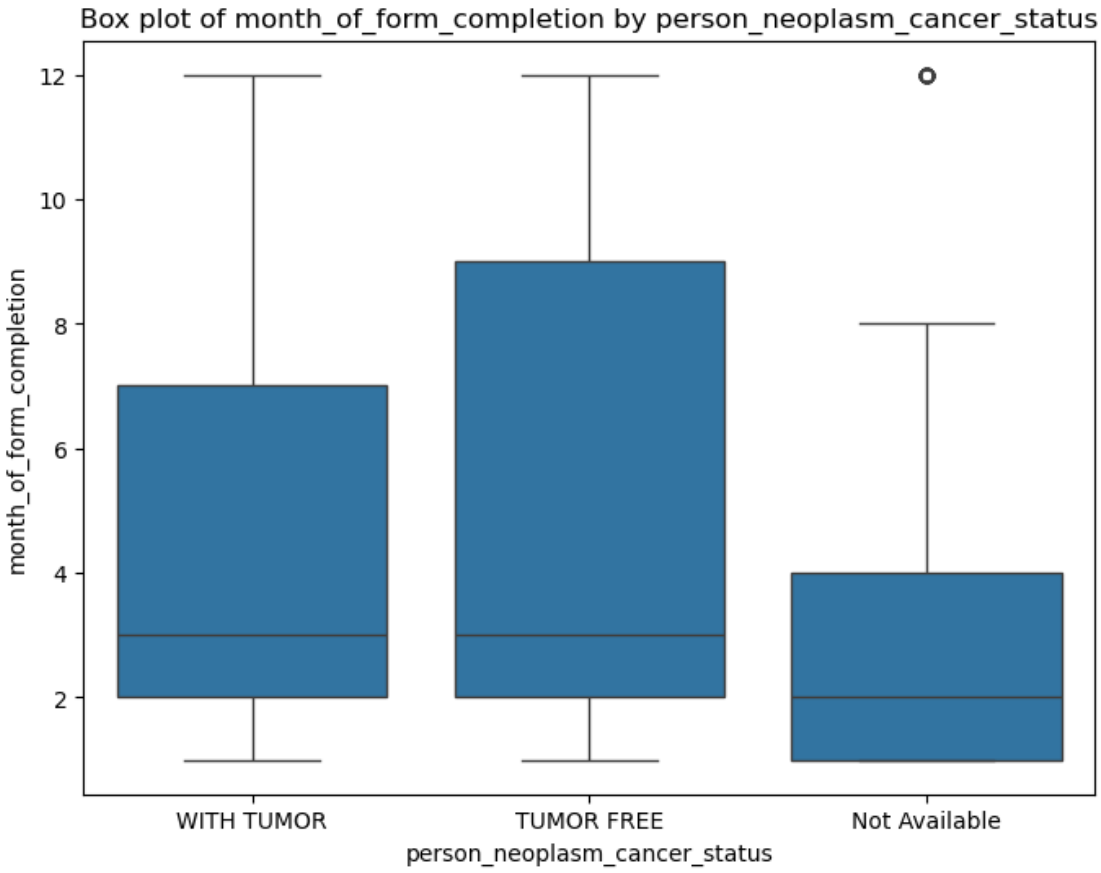

Box plot of weight by person_neoplasm_cancer_status
weight
200
180
160
140
120
100
80
60
40
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status

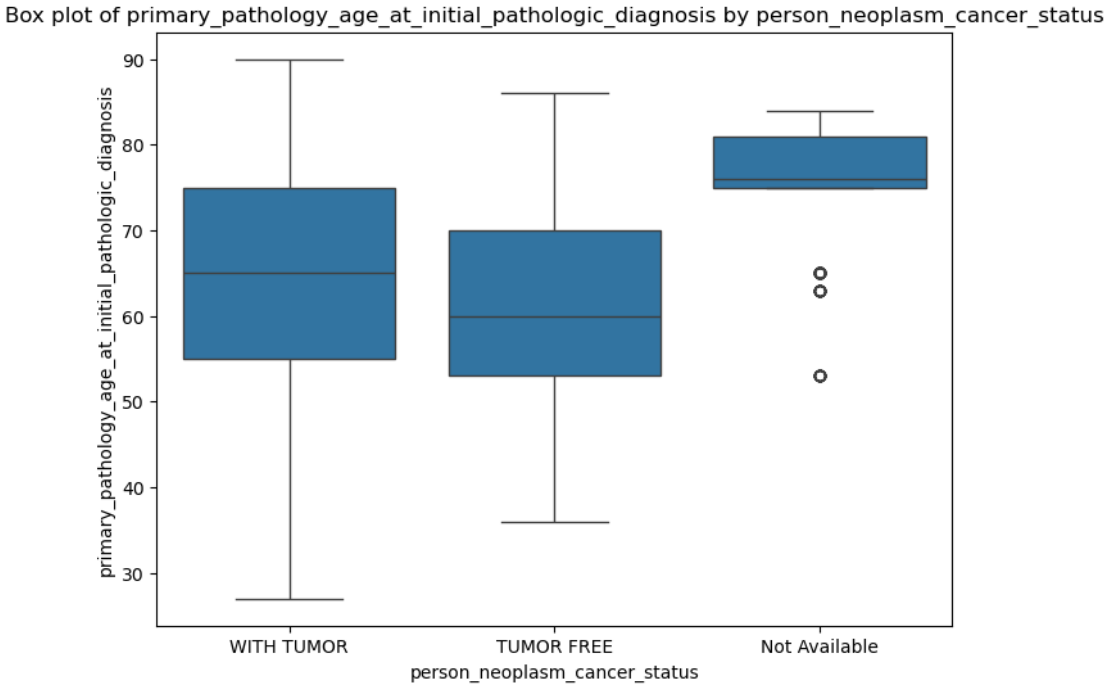

Box plot of days_to_last_followup by person_neoplasm_cancer_status
days_to_last_followup
3500
3000
2500
2000
1500
1000
500
0
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status


Box plot of number_pack_years_smoked by person_neoplasm_cancer_status
number_pack_years_smoked
0
20
40
60
80
100
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status

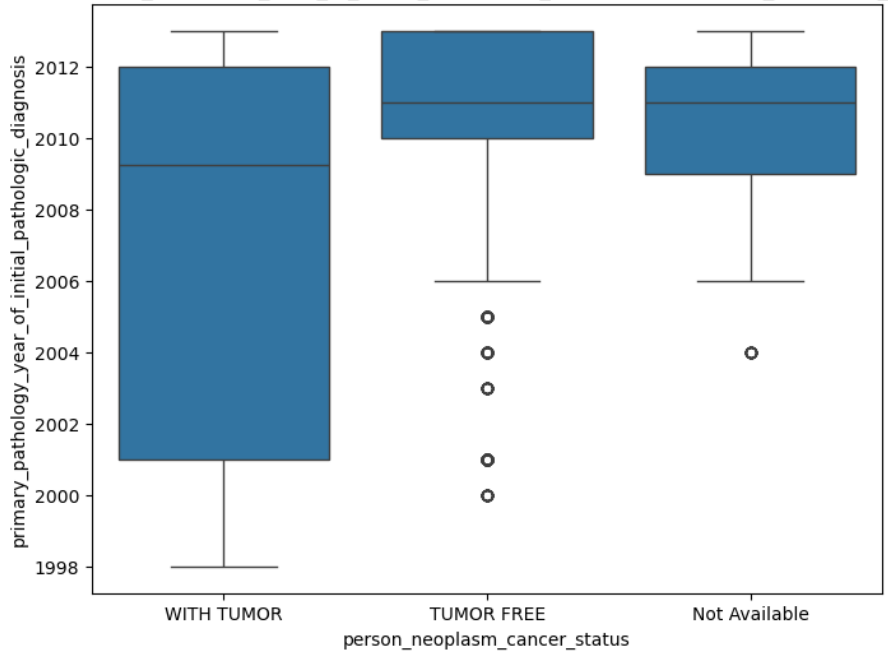

Box plot of amount_of_alcohol_consumption_per_day by person_neoplasm_cancer_status
amount_of_alcohol_consumption_per_day
14
12
10
8
6
4
2
0
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status


Box plot of day_of_form_completion by person_neoplasm_cancer_status
day_of_form_completion
30
25
20
15
10
5
0
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status

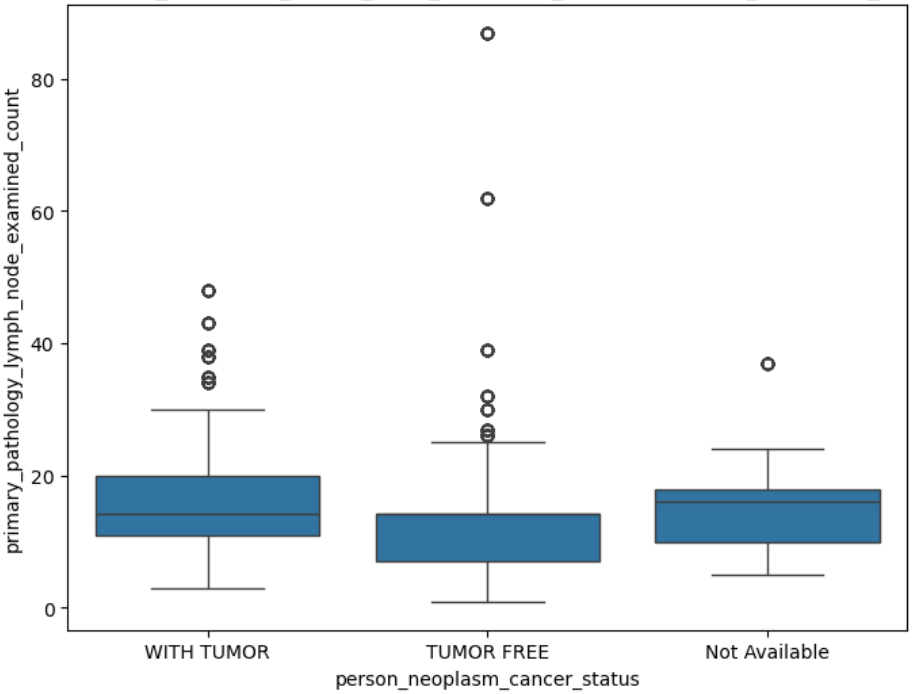

Box plot of month_of_form_completion by person_neoplasm_cancer_status
month_of_form_completion
12
10
8
6
4
2
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status

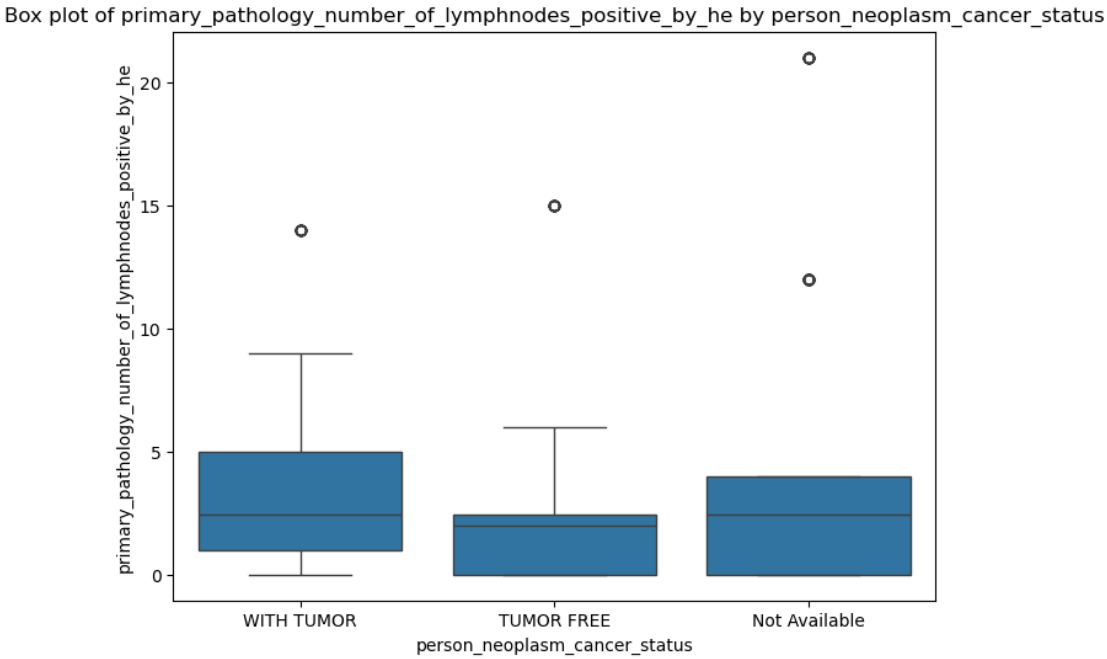

Box plot of primary_pathology_age_at_initial_pathologic_diagnosis by person_neoplasm_cancer_status
primary_pathology_age_at_initial_pathologic_diagnosis
90
80
70
60
50
40
30
WITH TUMOR
TUMOR FREE
Not Available
person_neoplasm_cancer_status

Box plot of primary_pathology_year_of_initial_pathologic_diagnosis by person_neoplasm_cancer_status

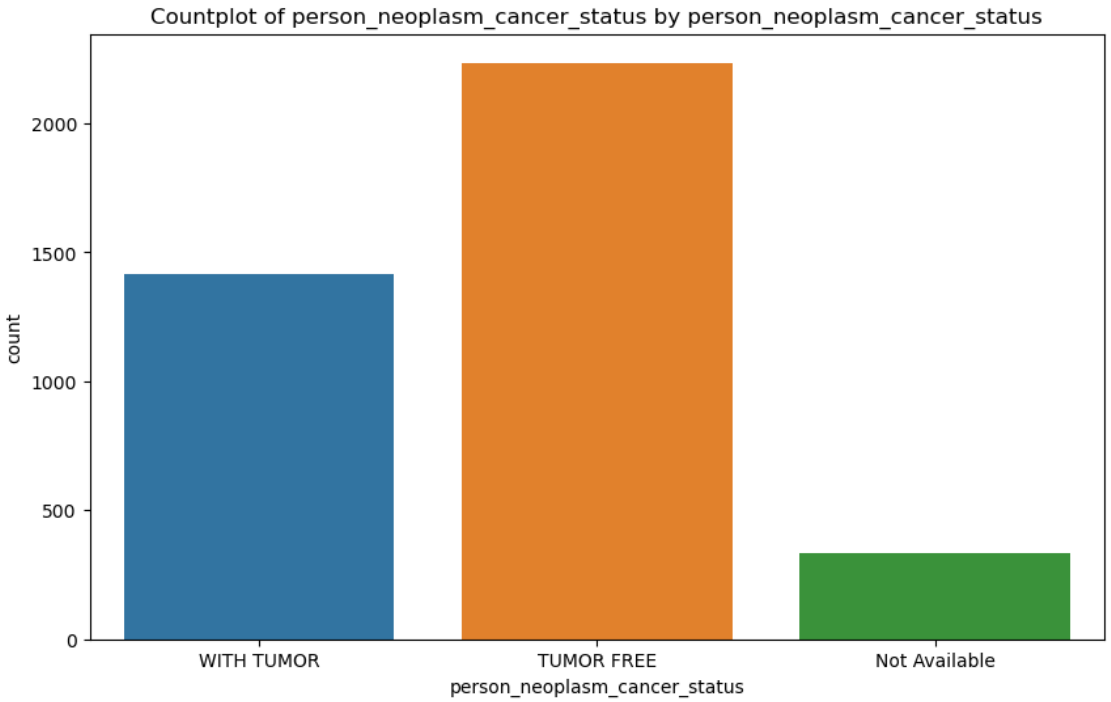


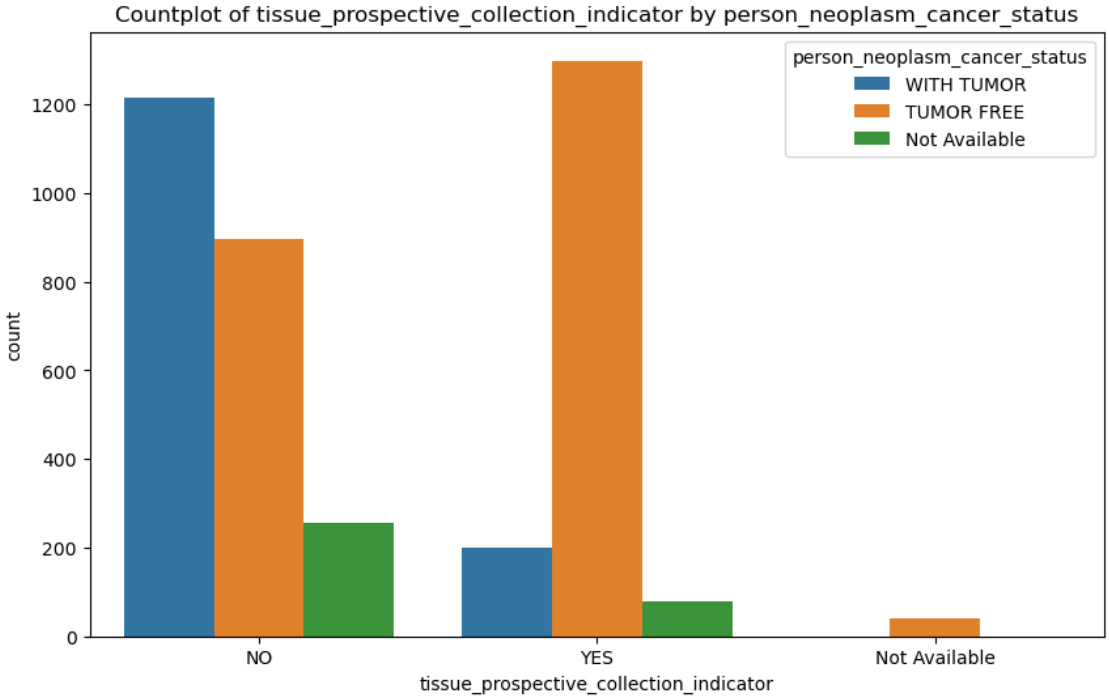
Box plot of primary_pathology_lymph_node_examined_count by person_neoplasm_cancer_status

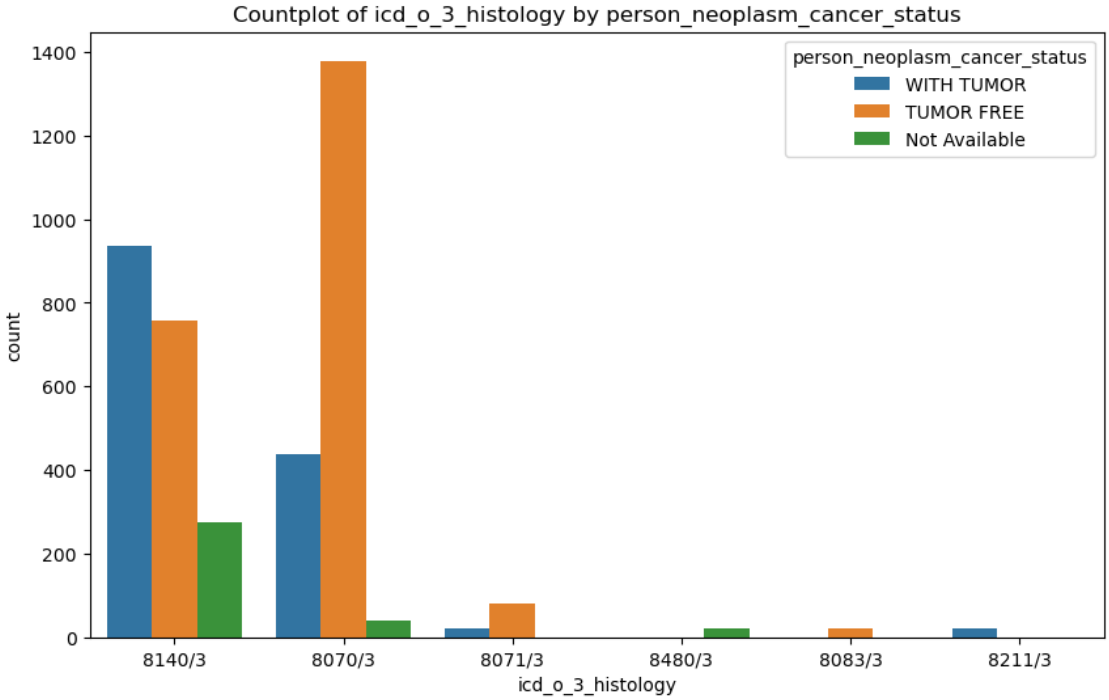


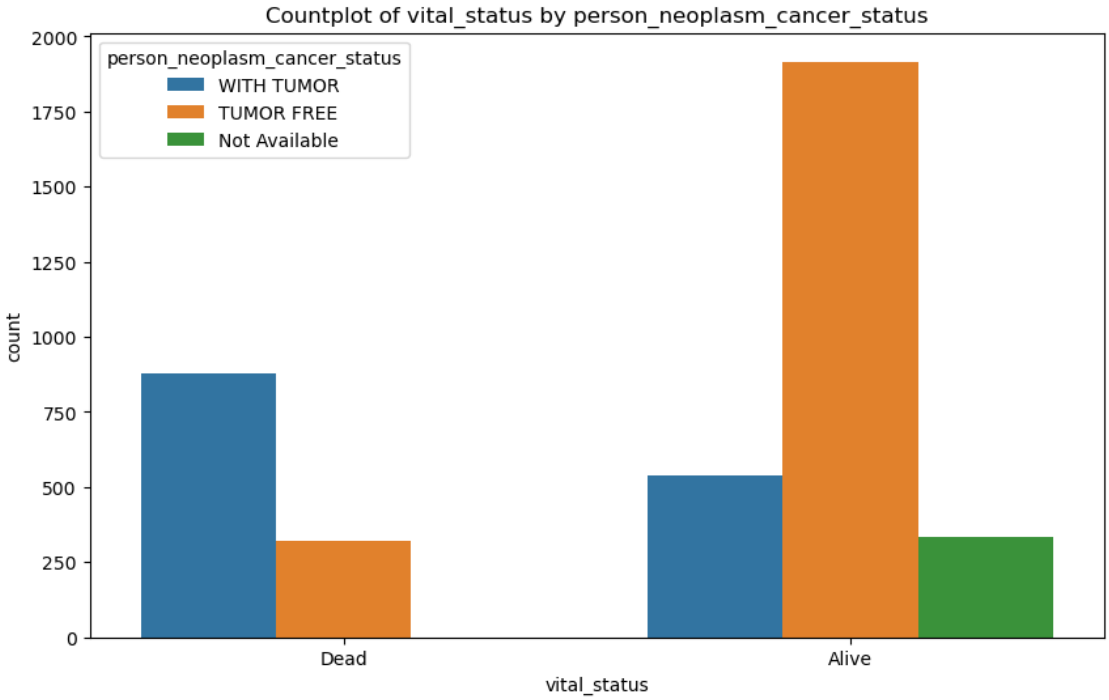


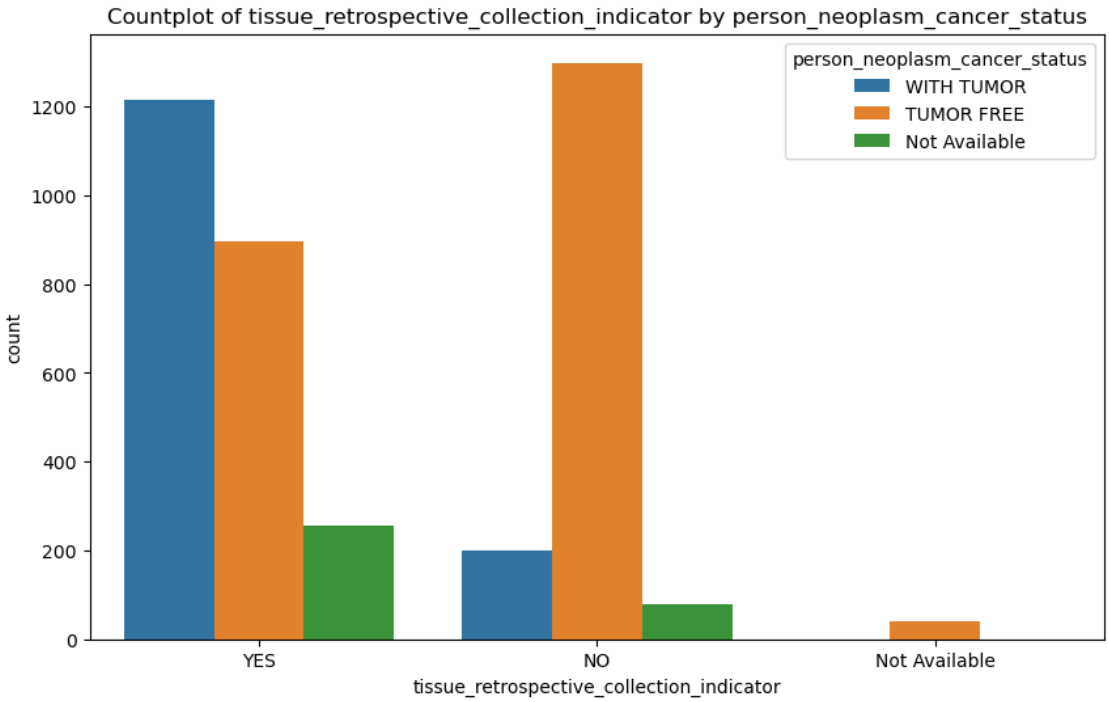
[42]:
```
for features in best_categorical_features:
    plt.figure(figsize=(10,6))
    sns.countplot(x=df[features], hue=df[target_variable])
    plt.title(f'Countplot of {features} by {target_variable}')
    plt.show()
```

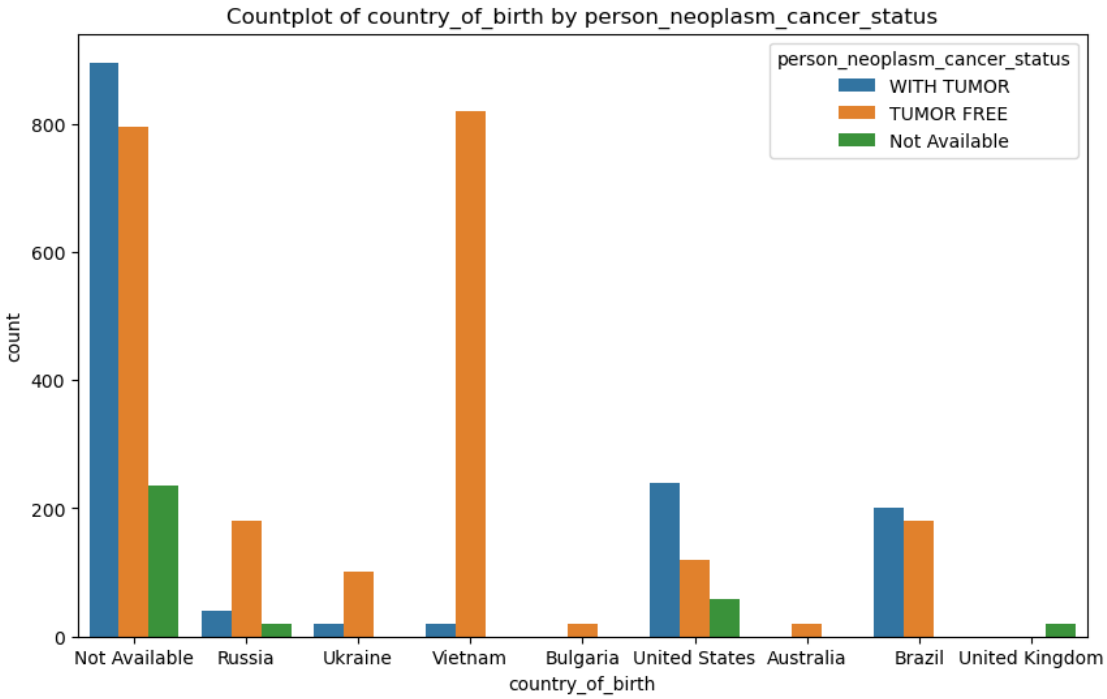


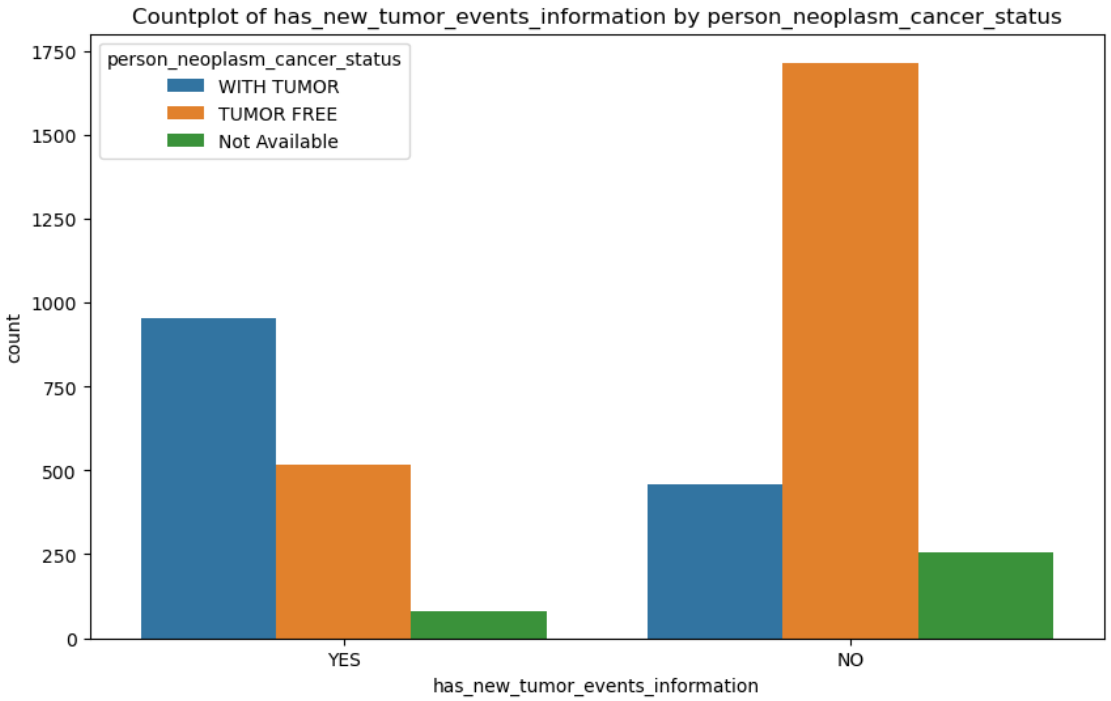

Countplot of tissue_prospective_collection_indicator by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
NO
YES
Not Available
tissue_prospective_collection_indicator

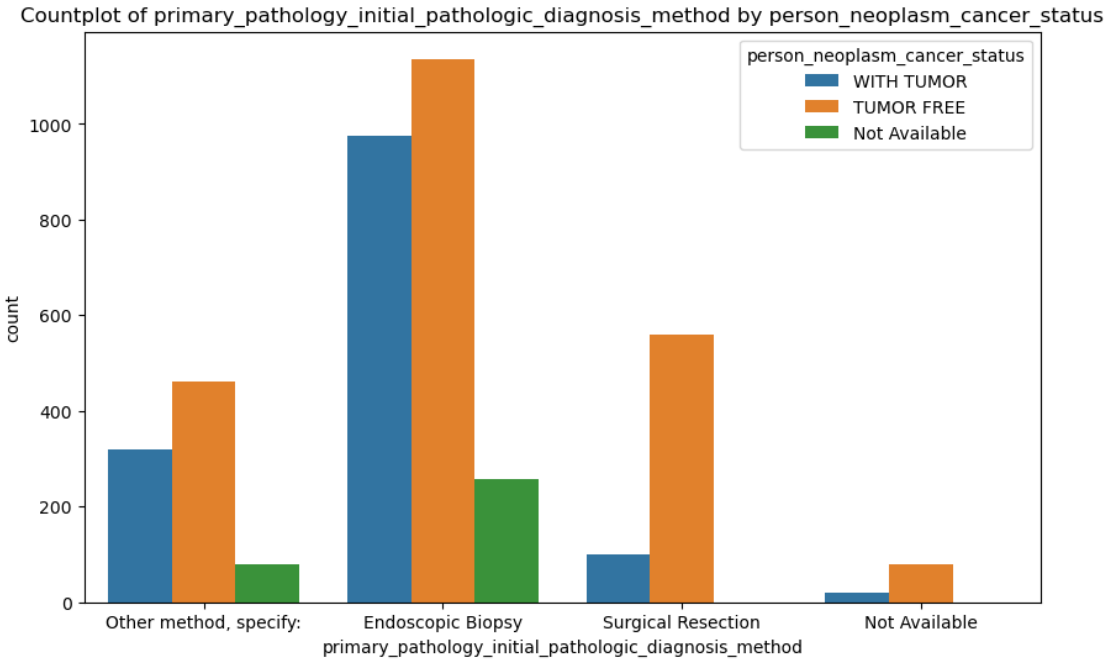

Countplot of icd_o_3_histology by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
8140/3
8070/3
8071/3
8480/3
8083/3
8211/3
icd_o_3_histology

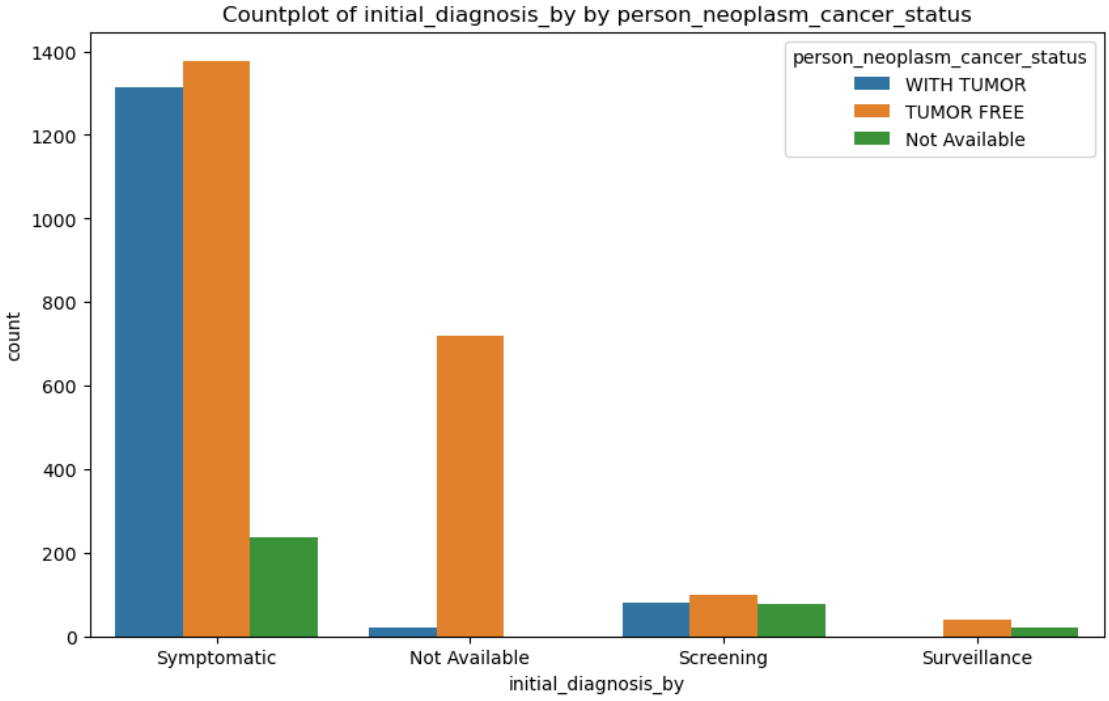

Countplot of vital_status by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
2000
1750
1500
1250
1000
750
500
250
0
Dead
Alive
vital_status

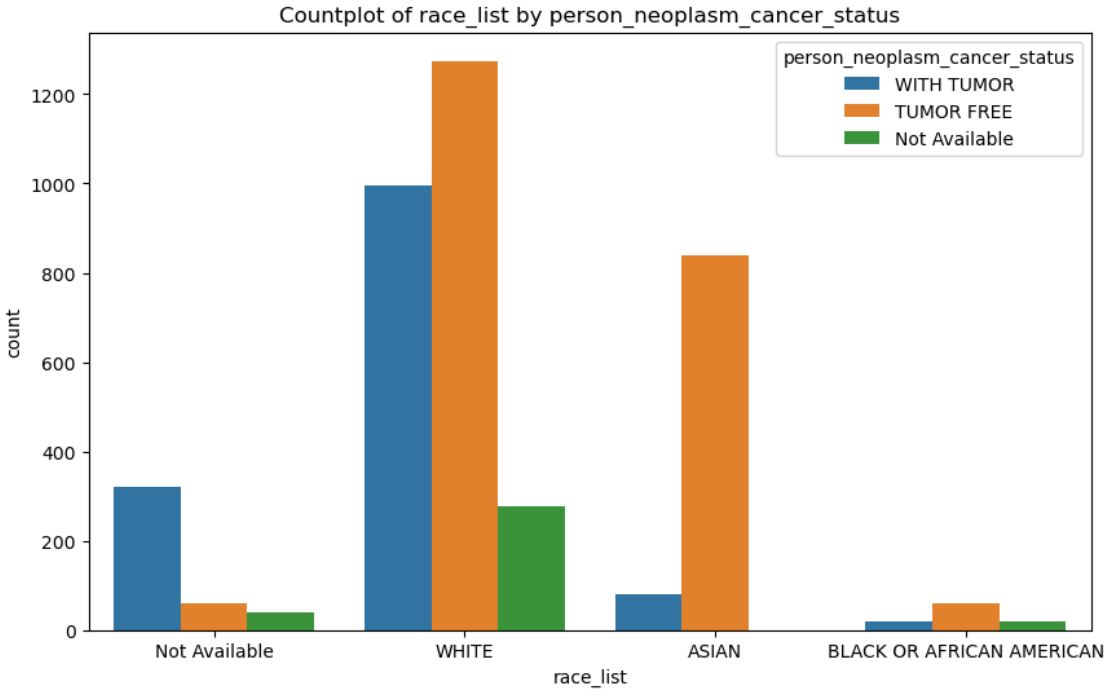

Countplot of tissue_retrospective_collection_indicator by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
1200
1000
800
600
400
200
0
YES
NO
Not Available
tissue_retrospective_collection_indicator

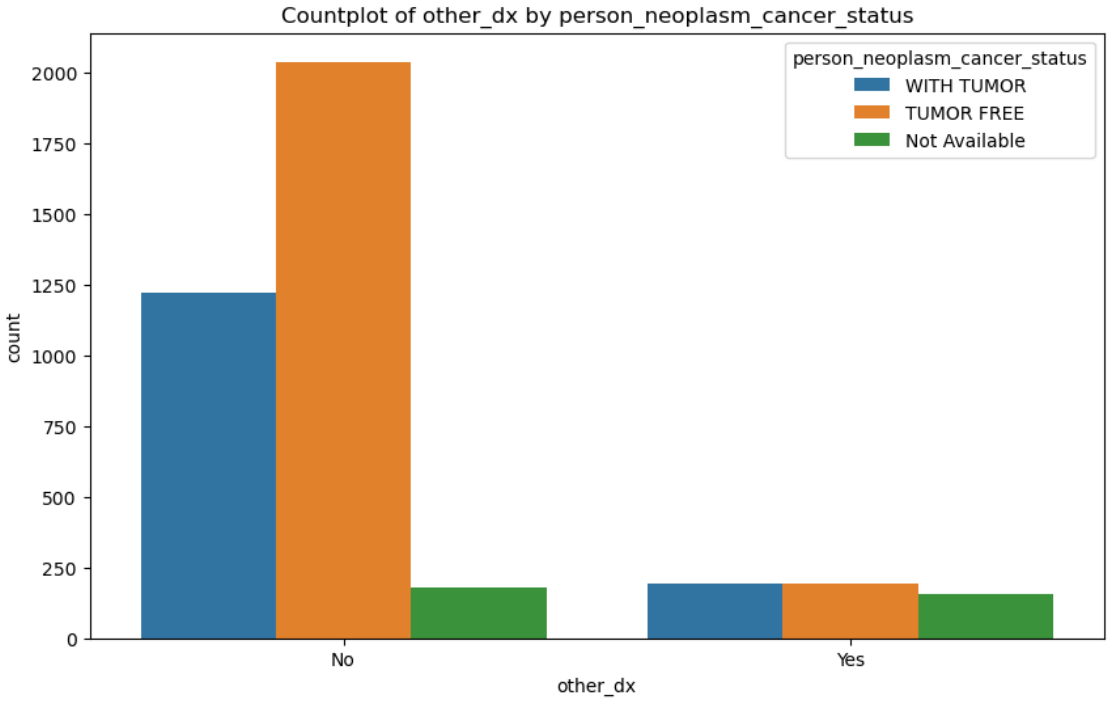

Countplot of country_of_birth by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
Not Available
Russia
Ukraine
Vietnam
Bulgaria
United States
Australia
Brazil
United Kingdom
country_of_birth

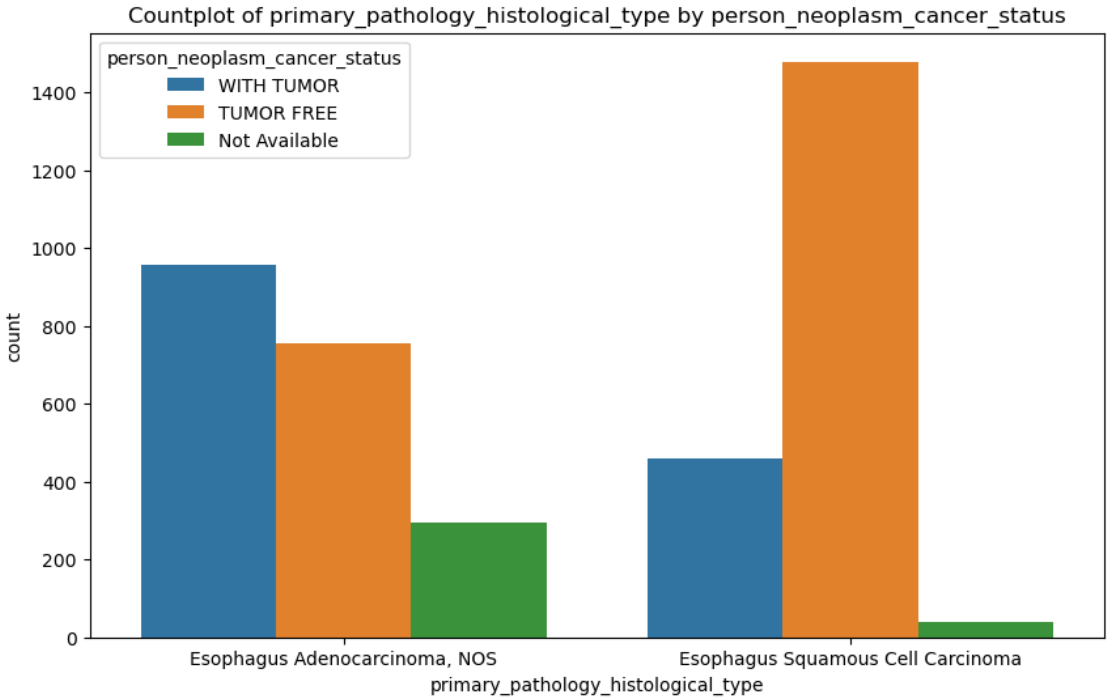

Countplot of has_new_tumor_events_information by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
250
500
750
1000
1250
1500
1750
YES
NO
has_new_tumor_events_information

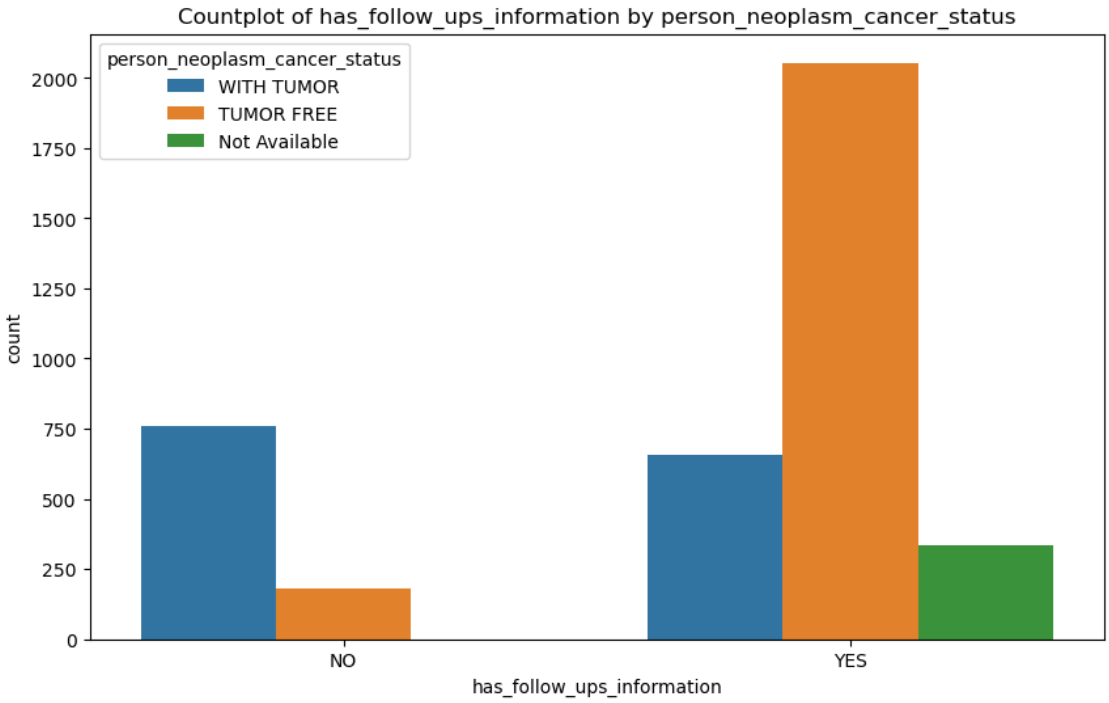

Countplot of primary_pathology_initial_pathologic_diagnosis_method by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
Other method, specify:
Endoscopic Biopsy
Surgical Resection
Not Available
primary_pathology_initial_pathologic_diagnosis_method

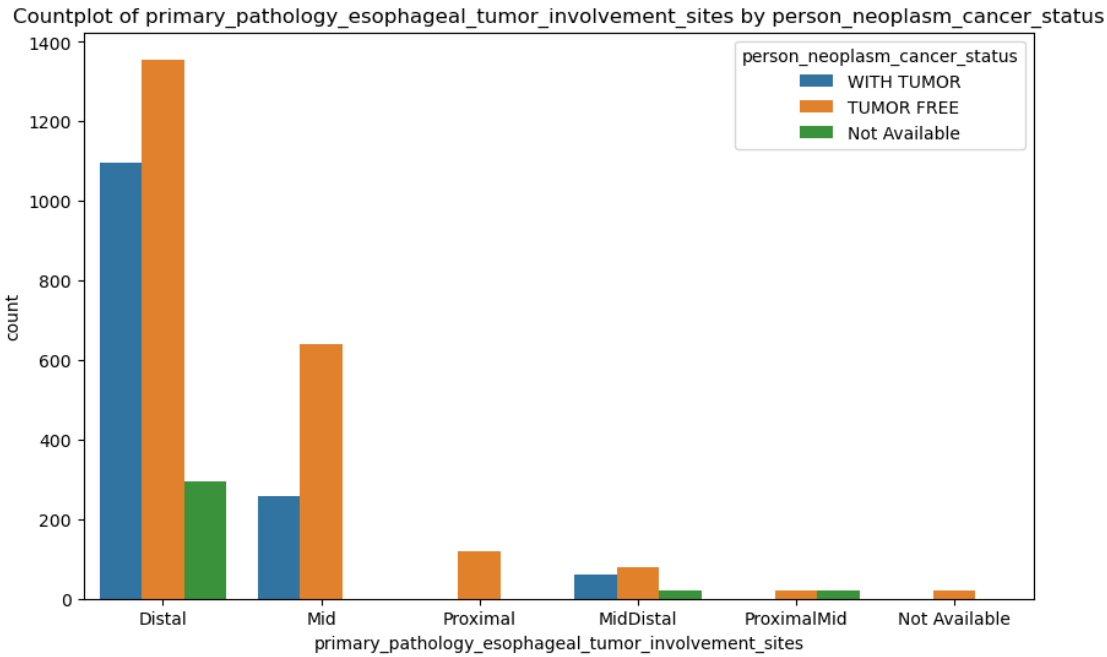

Countplot of initial_diagnosis_by by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
Symptomatic
Not Available
Screening
Surveillance
initial_diagnosis_by

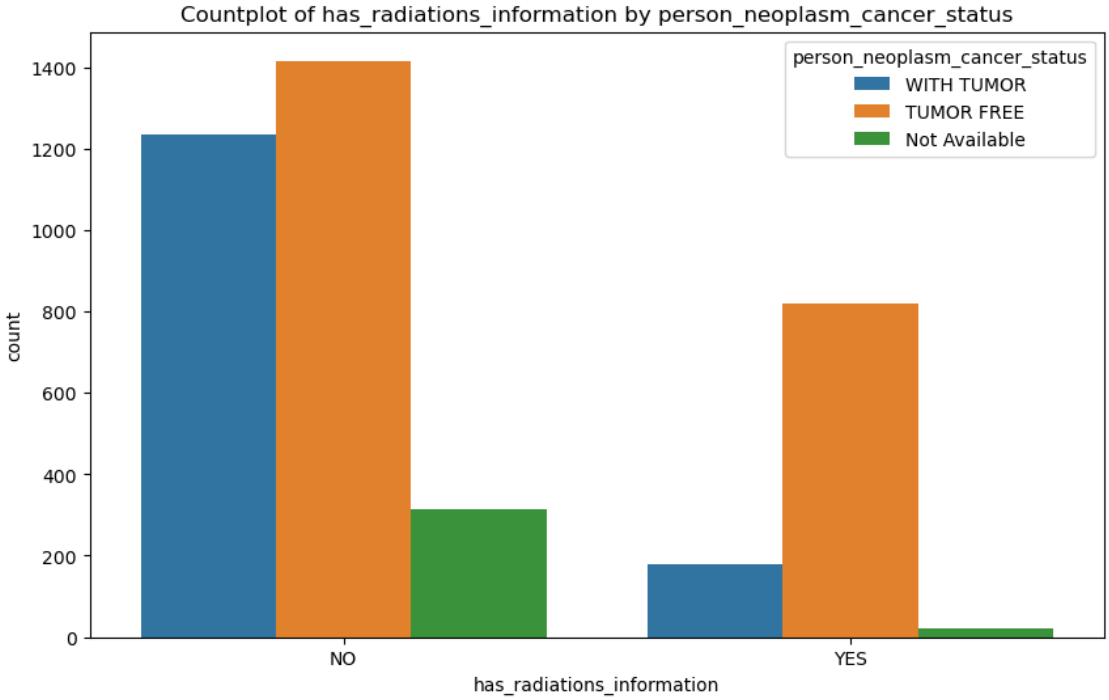

Countplot of race_list by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
Not Available
WHITE
ASIAN
BLACK OR AFRICAN AMERICAN
race_list

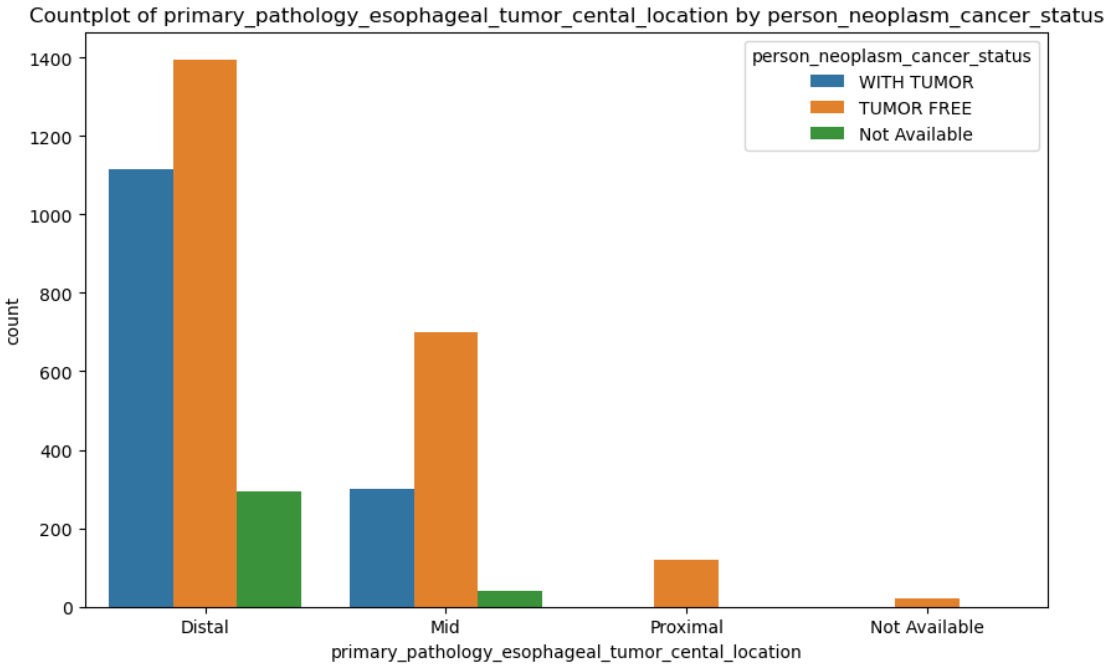

Countplot of other_dx by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
250
500
750
1000
1250
1500
1750
2000
No
Yes
other_dx

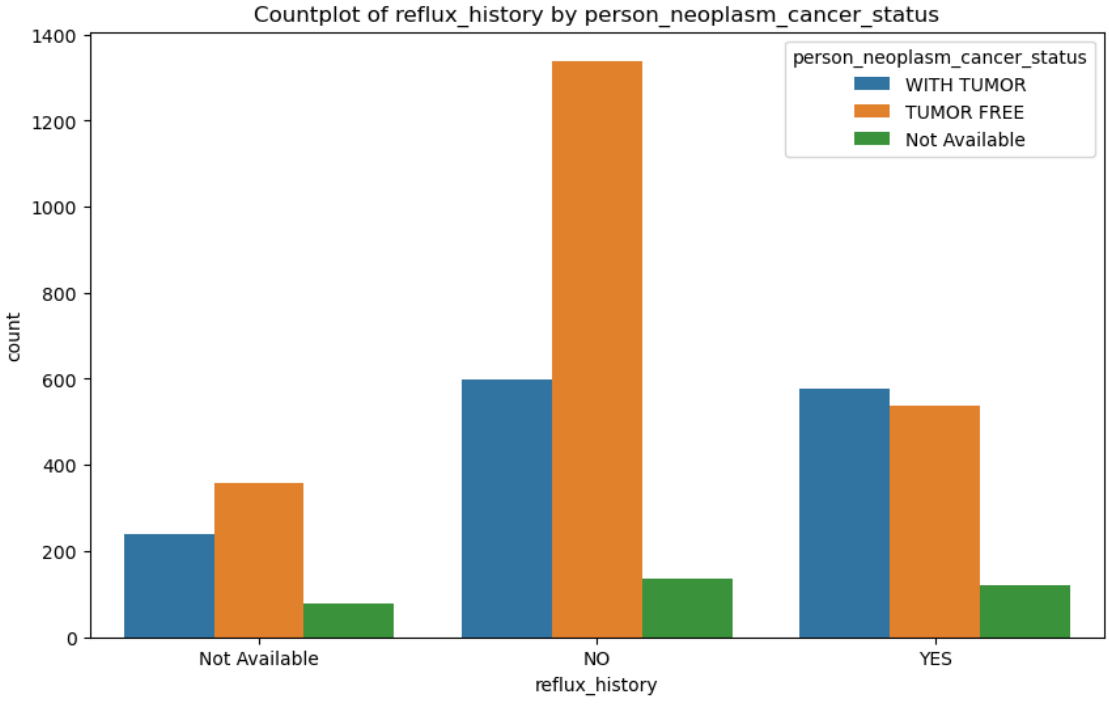

Countplot of primary_pathology_histological_type by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
Esophagus Adenocarcinoma, NOS
Esophagus Squamous Cell Carcinoma
primary_pathology_histological_type

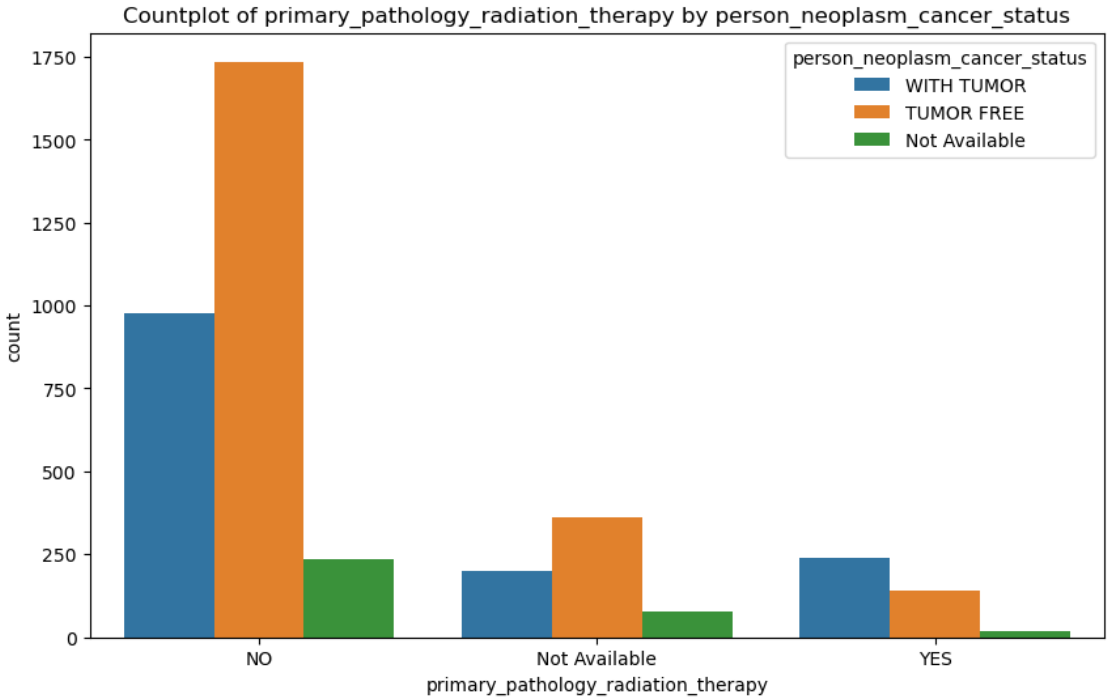

Countplot of has_follow_ups_information by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
250
500
750
1000
1250
1500
1750
2000
NO
YES
has_follow_ups_information

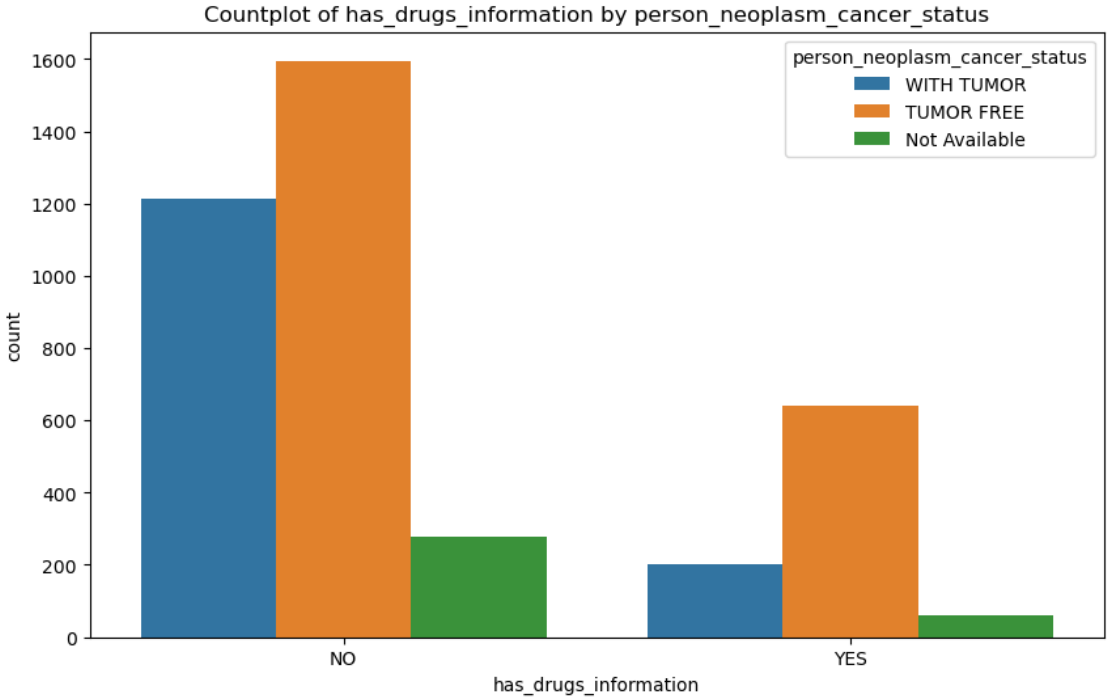

Countplot of primary_pathology_esophageal_tumor_involvement_sites by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
Distal
Mid
Proximal
MidDistal
ProximalMid
Not Available
primary_pathology_esophageal_tumor_involvement_sites

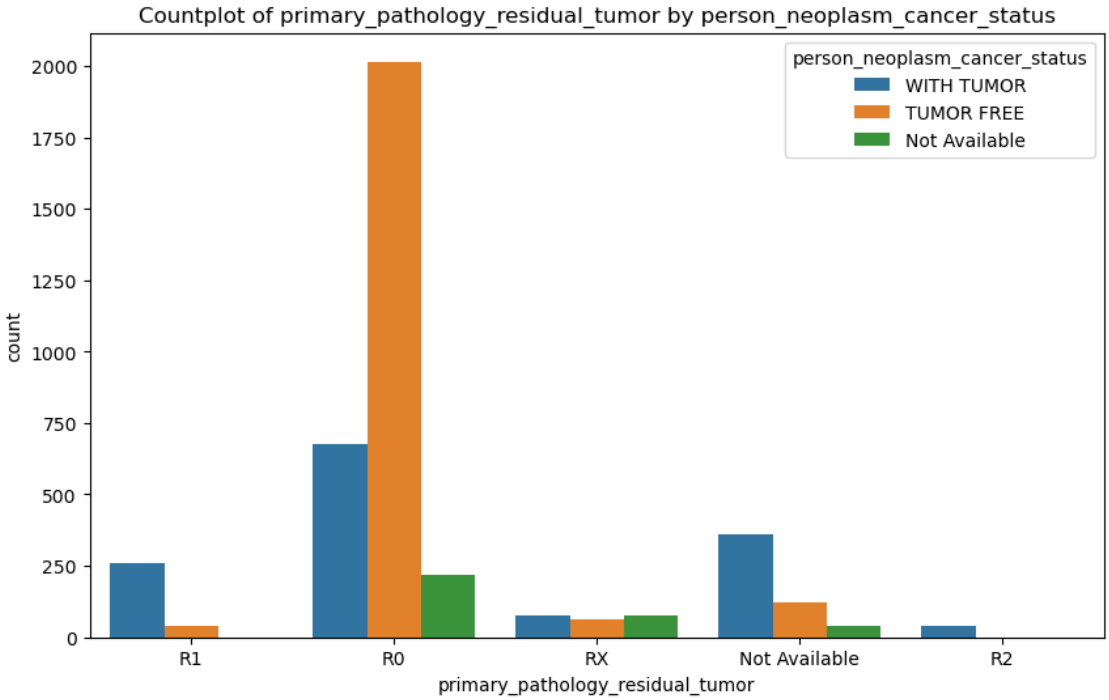

Countplot of has_radiations_information by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
NO
YES
has_radiations_information

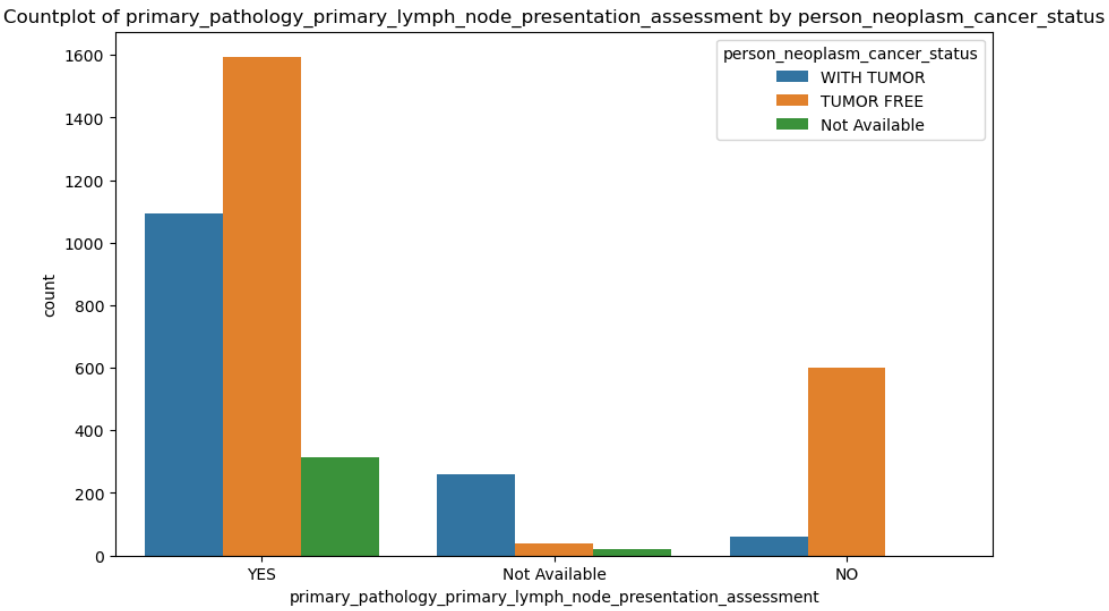

Countplot of primary_pathology_esophageal_tumor_cental_location by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
Distal
Mid
Proximal
Not Available
primary_pathology_esophageal_tumor_cental_location

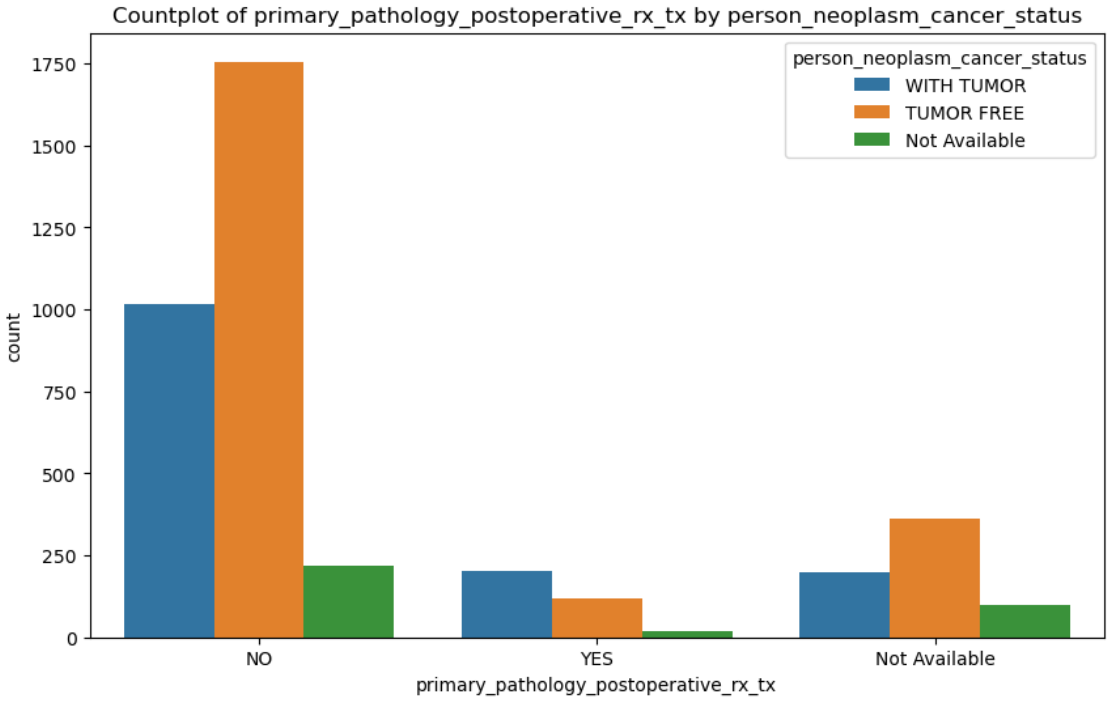

Countplot of reflux_history by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
Not Available
NO
YES
reflux_history

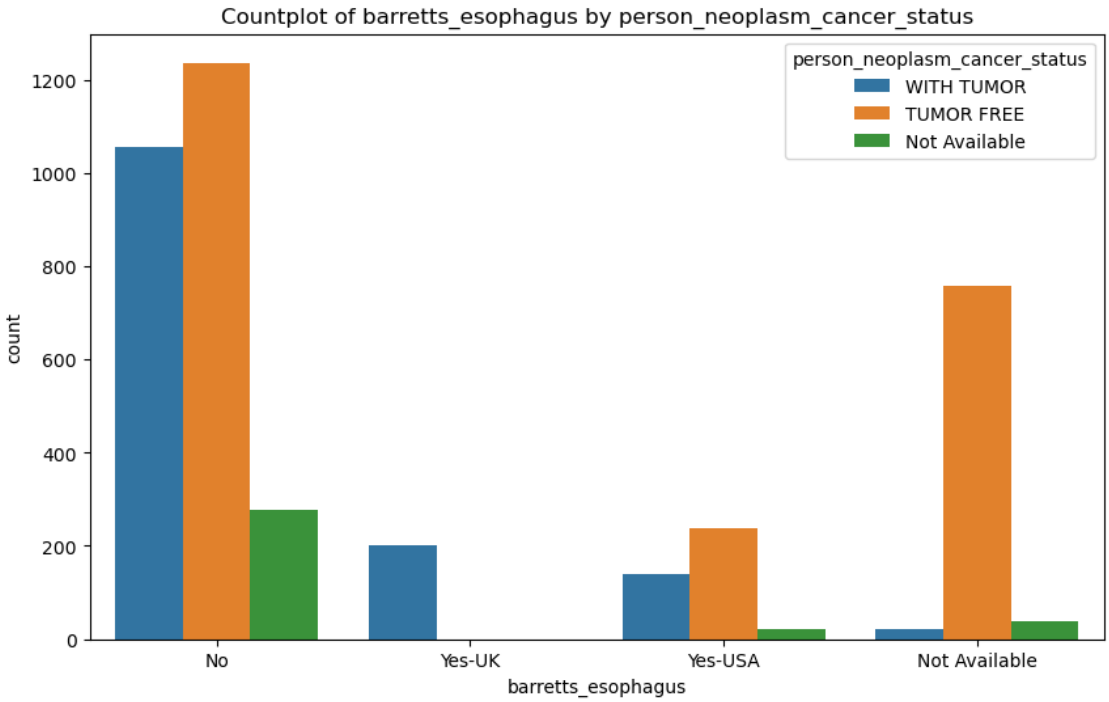

Countplot of primary_pathology_radiation_therapy by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
1750
1500
1250
1000
750
500
250
0
NO
Not Available
YES
primary_pathology_radiation_therapy

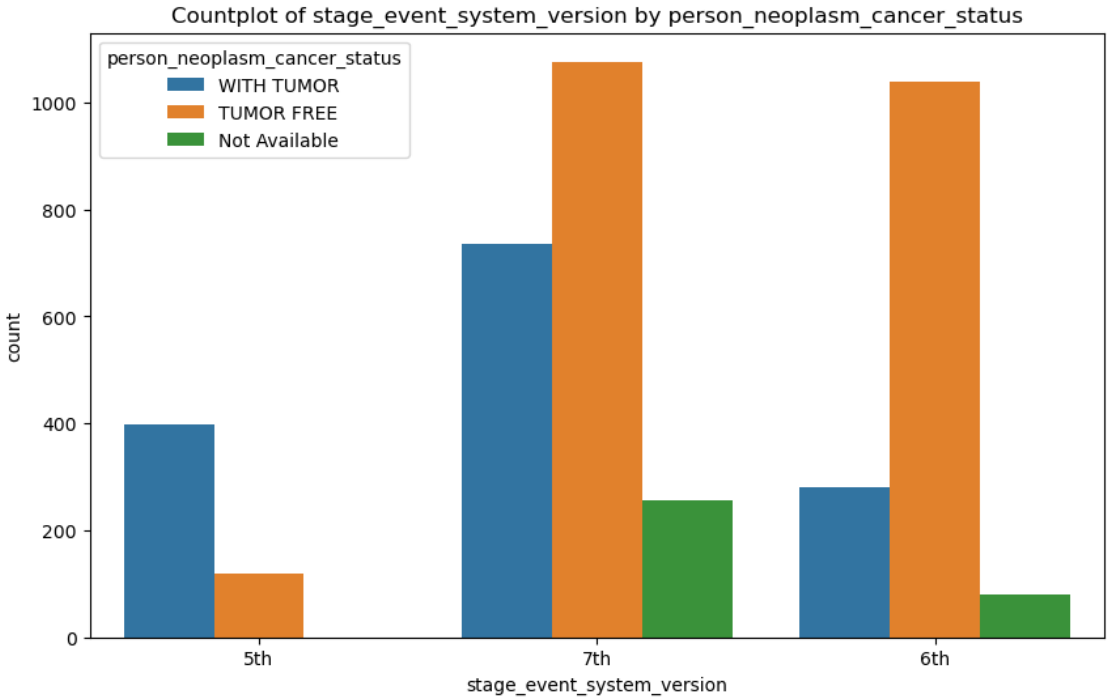

Countplot of has_drugs_information by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
1600
1400
1200
1000
800
600
400
200
0
NO
YES
has_drugs_information

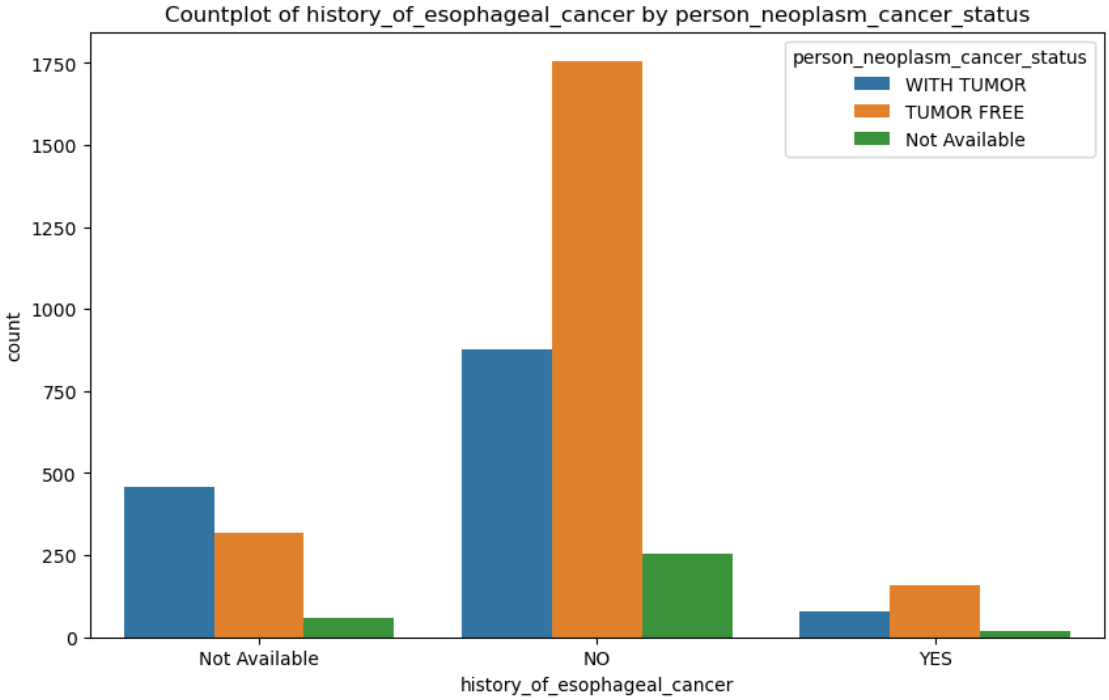

Countplot of primary_pathology_residual_tumor by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
2000
1750
1500
1250
1000
750
500
250
0
R1
R0
RX
Not Available
R2
primary_pathology_residual_tumor

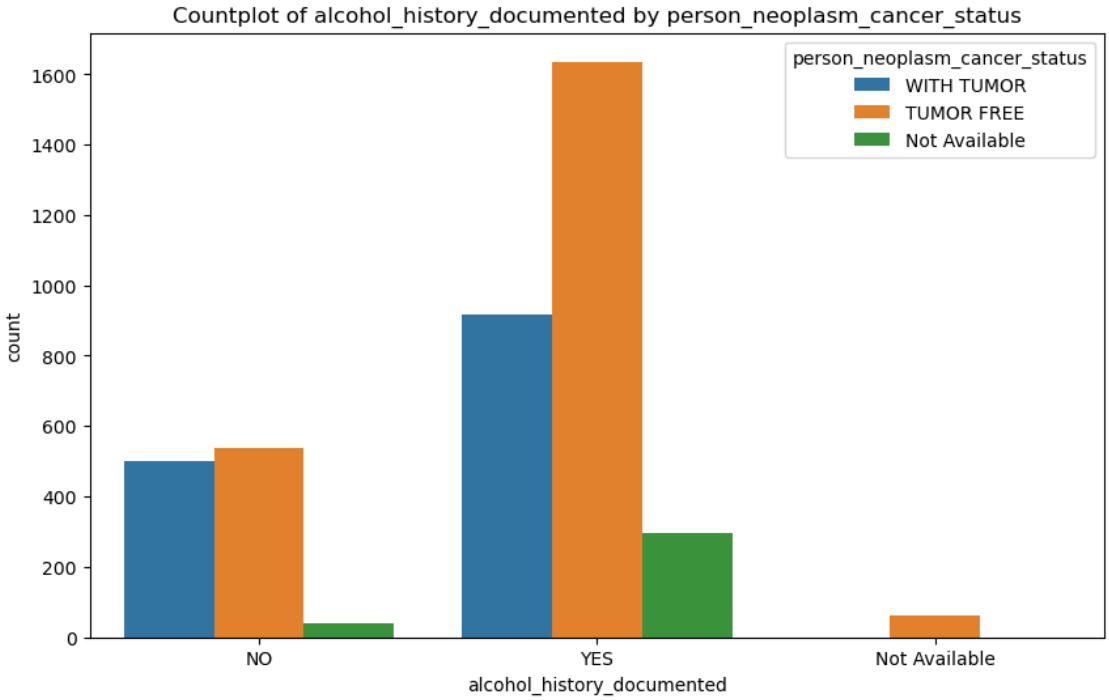

Countplot of primary_pathology_primary_lymph_node_presentation_assessment by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
1600
1400
1200
1000
800
600
400
200
0
YES
Not Available
NO
primary_pathology_primary_lymph_node_presentation_assessment

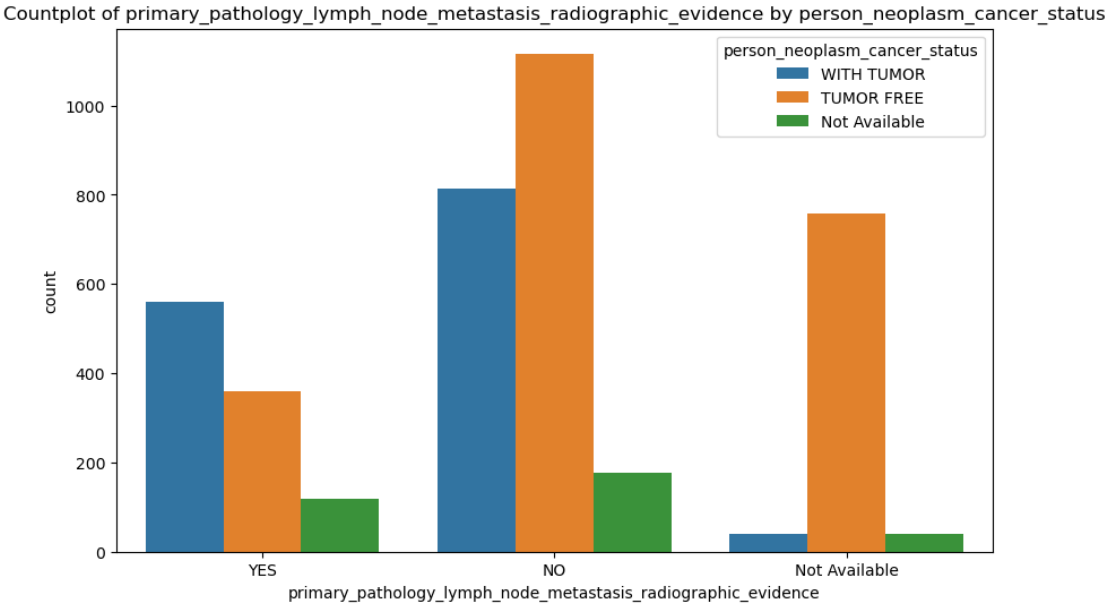

Countplot of primary_pathology_postoperative_rx_tx by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
1750
1500
1250
1000
750
500
250
0
NO
YES
Not Available
primary_pathology_postoperative_rx_tx

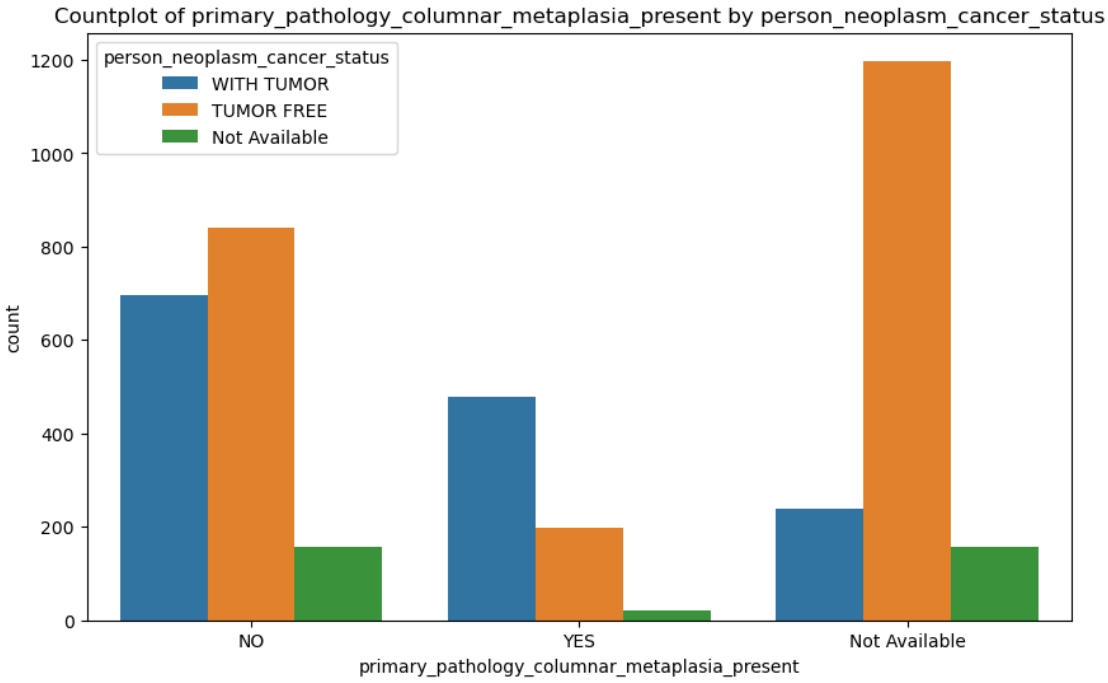

Countplot of barretts_esophagus by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
1200
1000
800
600
400
200
0
No
Yes-UK
Yes-USA
Not Available
barretts_esophagus

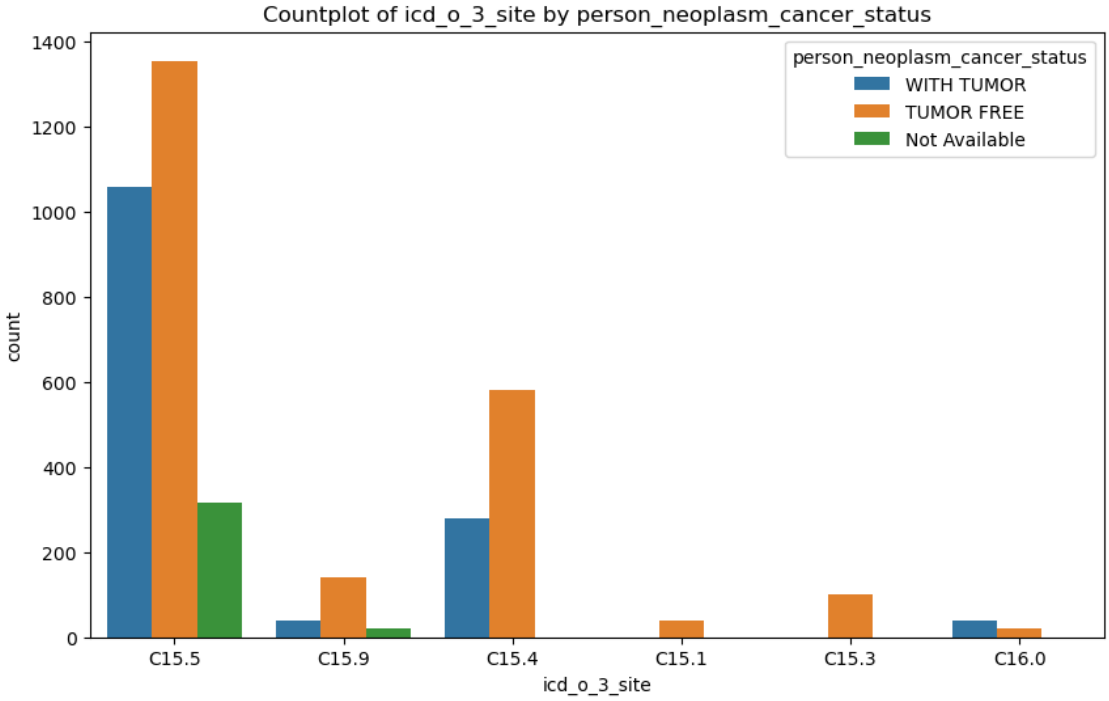

Countplot of stage_event_system_version by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
5th
7th
6th
stage_event_system_version

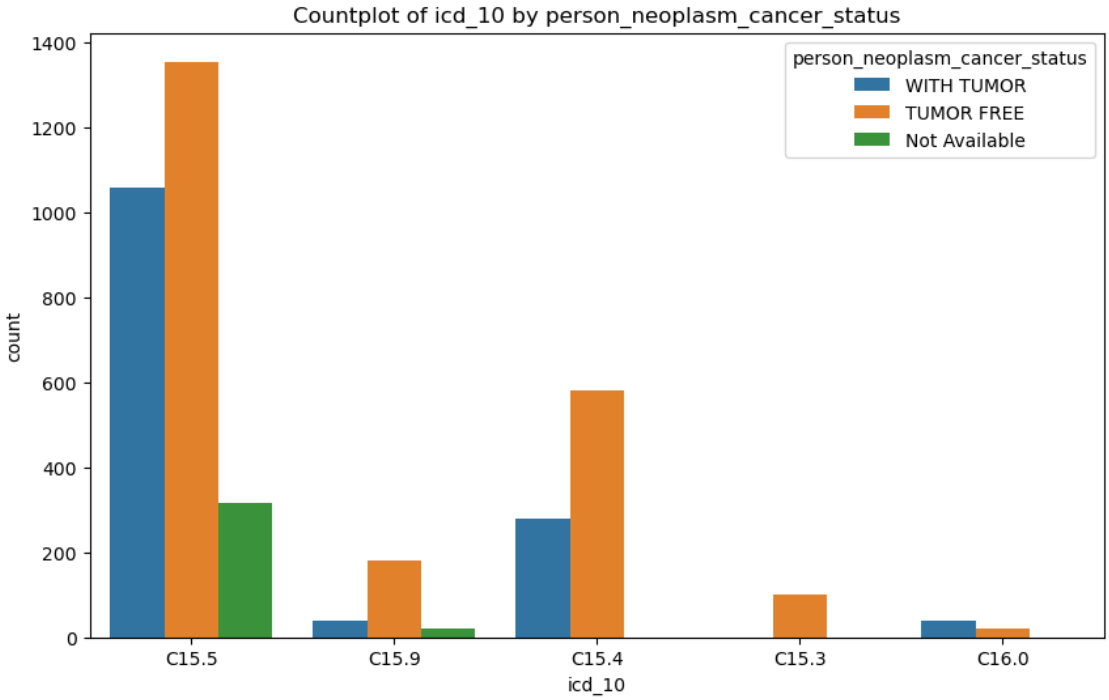

Countplot of history_of_esophageal_cancer by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
250
500
750
1000
1250
1500
1750
Not Available
NO
YES
history_of_esophageal_cancer

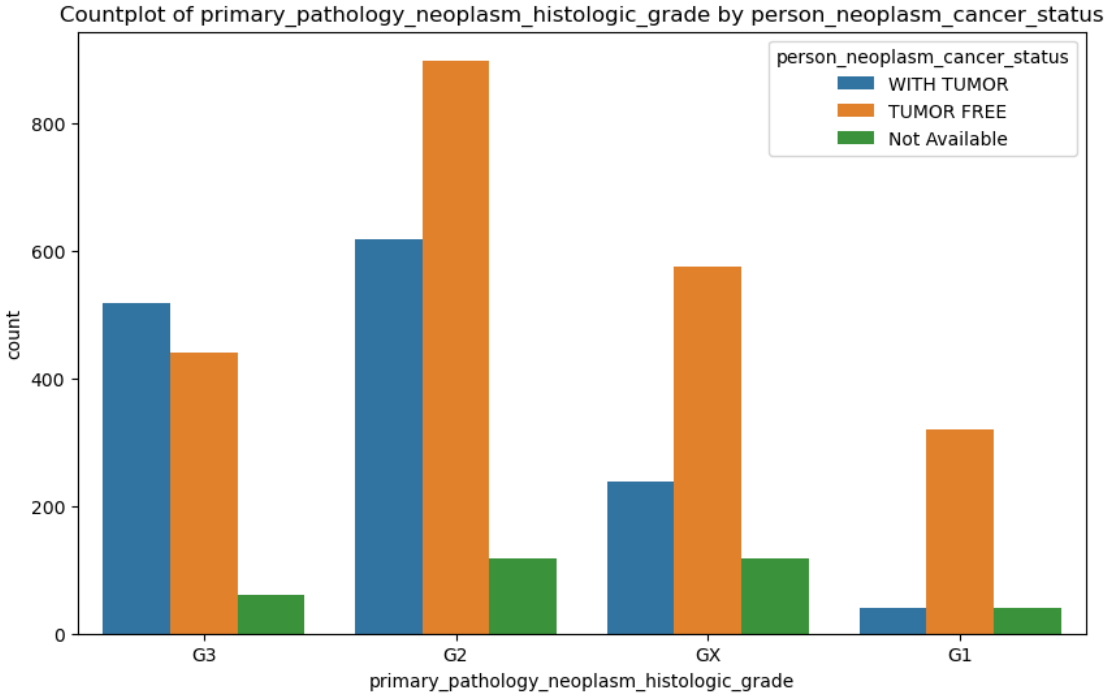

Countplot of alcohol_history_documented by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
1600
NO
YES
Not Available
alcohol_history_documented


Countplot of primary_pathology_lymph_node_metastasis_radiographic_evidence by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
YES
NO
Not Available
primary_pathology_lymph_node_metastasis_radiographic_evidence

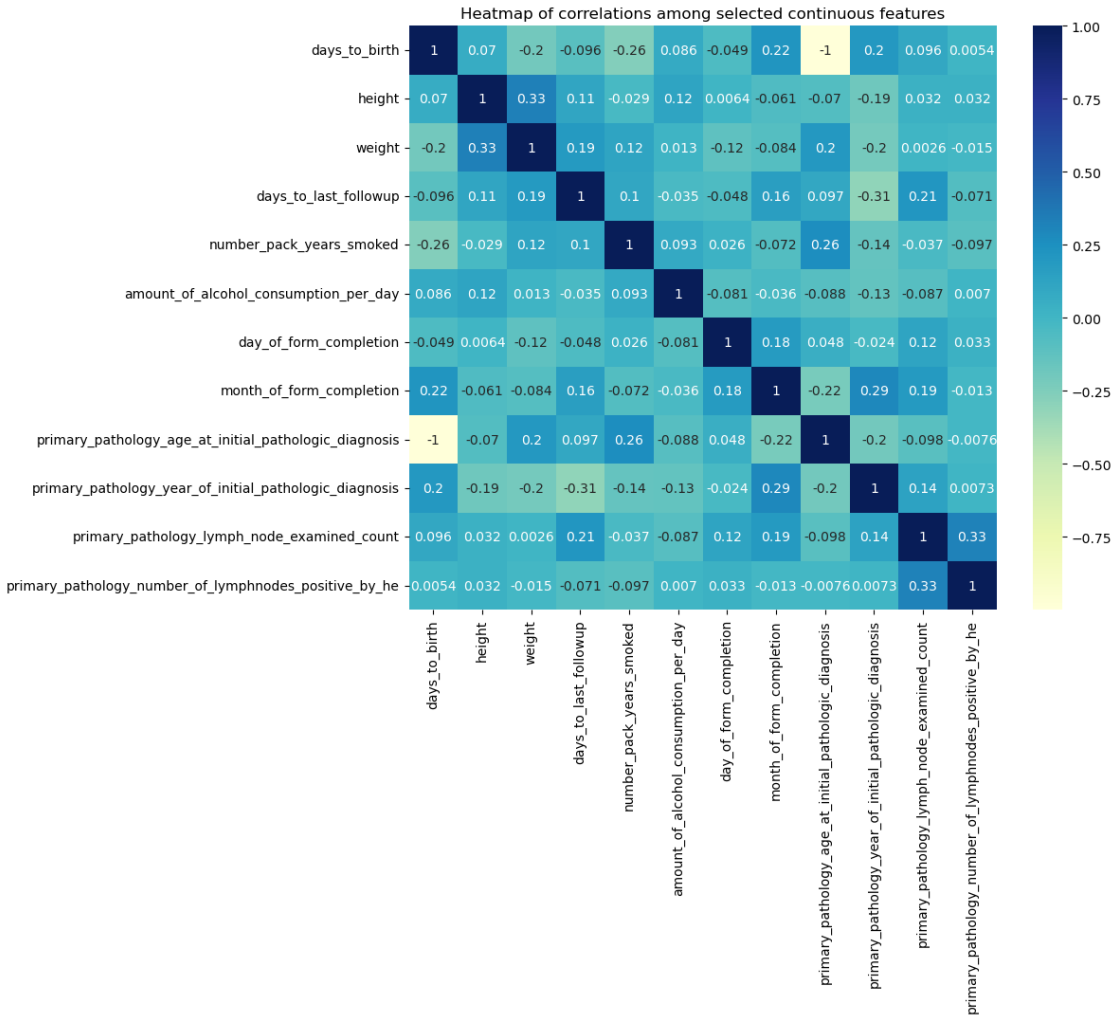

Countplot of primary_pathology_columnar_metaplasia_present by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
NO
YES
Not Available
primary_pathology_columnar_metaplasia_present


Countplot of icd_o_3_site by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
C15.5
C15.9
C15.4
C15.1
C15.3
C16.0
icd_o_3_site


Countplot of icd_10 by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
1000
1200
1400
C15.5
C15.9
C15.4
C15.3
C16.0
icd_10


Countplot of primary_pathology_neoplasm_histologic_grade by person_neoplasm_cancer_status
person_neoplasm_cancer_status
WITH TUMOR
TUMOR FREE
Not Available
count
0
200
400
600
800
G3
G2
GX
G1
primary_pathology_neoplasm_histologic_grade

[44]:
```
correlation_matrix = df[continuous_features].corr()
plt.figure(figsize=(10,8))
sns.heatmap(correlation_matrix, annot = True, cmap="YlGnBu")
plt.title("Heatmap of correlations among selected continuous features")
plt.show()
```



[45]:
```
pivot_table_mean = df.pivot_table(index = 'gender', columns='race_list',␣
↪values='days_to_birth', aggfunc='mean')
pivot_table_mean
```

[45]:
```
race_list          ASIAN  BLACK OR AFRICAN AMERICAN  Not Available  \
gender
FEMALE    -21819.883333                        NaN  -27982.556962
MALE      -20460.469767                  -21408.26  -22807.420588
```

```
race_list         WHITE
gender
FEMALE    -24438.048218
MALE      -24384.117931
```

[46]:
```
pivot_table_max_min = df.pivot_table(index = 'gender',␣
 ↪columns='person_neoplasm_cancer_status',
                                     ␣
 ↪values='primary_pathology_age_at_initial_pathologic_diagnosis',
                                      aggfunc=['max','min']
)
pivot_table_max_min
```

[46]:
```
                                        max                        \
person_neoplasm_cancer_status Not Available TUMOR FREE WITH TUMOR
gender
FEMALE                                 84         86         84
MALE                                   81         86         90

                                        min
person_neoplasm_cancer_status Not Available TUMOR FREE WITH TUMOR
gender
FEMALE                                 79         44         51
MALE                                   53         36         27
```

[47]:
```
pivot_table_multi_agg = df.pivot_table(
    index='primary_pathology_histological_type',
    columns='vital_status',
    values='days_to_last_followup',
    aggfunc=['mean','median','std']
)
pivot_table_multi_agg
```

[47]:
```
                                          mean             median  \
vital_status                             Alive        Dead  Alive
primary_pathology_histological_type
Esophagus Adenocarcinoma, NOS       530.138324  306.201937  267.0
Esophagus Squamous Cell Carcinoma   129.667094  306.201937    3.0

                                                    std
vital_status                              Dead       Alive Dead
primary_pathology_histological_type
Esophagus Adenocarcinoma, NOS       306.201937  619.523167  0.0
Esophagus Squamous Cell Carcinoma   306.201937  291.660826  0.0
```

[48]:
```
pivot_table_percentage = df.pivot_table(
    index='gender',
```

```
    columns='person_neoplasm_cancer_status',
    values='patient_id',
    aggfunc='count'
)
pivot_table_percentage = pivot_table_percentage.div(pivot_table_percentage.
 ↪sum(axis=1), axis=0)*100
pivot_table_percentage
```

[48]:
```
person_neoplasm_cancer_status  Not Available  TUMOR FREE  WITH TUMOR
gender
FEMALE                             12.662338   64.772727   22.564935
MALE                                7.628376   54.496883   37.874740
```

[50]:
```
pivot_table_totals = df.pivot_table(
    index='tissue_prospective_collection_indicator',
    columns='country_of_birth',
    values='primary_pathology_age_at_initial_pathologic_diagnosis',
    aggfunc='mean',
    margins=True,
    margins_name='Total'
)
pivot_table_totals
```

[50]:
```
country_of_birth                          Australia     Brazil  Bulgaria  \
tissue_prospective_collection_indicator
NO                                             72.0  56.411765       NaN
Not Available                                   NaN  58.500000       NaN
YES                                             NaN        NaN      50.0
Total                                          72.0  56.631579      50.0

country_of_birth                          Not Available  Russia    Ukraine  \
tissue_prospective_collection_indicator
NO                                            69.594156     NaN        NaN
Not Available                                       NaN     NaN        NaN
YES                                           64.696203   58.75  59.833333
Total                                         69.393358   58.75  59.833333

country_of_birth                          United Kingdom  United States  \
tissue_prospective_collection_indicator
NO                                                  65.0      54.857143
Not Available                                        NaN            NaN
YES                                                  NaN      64.053957
Total                                               65.0      60.973684

country_of_birth                            Vietnam      Total
tissue_prospective_collection_indicator
NO                                              NaN  66.811655
```

```
Not Available                                  NaN  58.500000
YES                                      56.214286  58.603678
Total                                    56.214286  63.480050
```

[70]:
```
correlation_matric = df[continuous_features].corr().abs()
threshold = 0.65
features_to_drop = set()
for i in range(len(correlation_matrix.columns)):
    for j in range(i):
        if correlation_matrix.iloc[i, j] > threshold:
            colname = correlation_matrix.columns[j]
            features_to_drop.add(colname)

df_reduced = df.drop(columns = features_to_drop)
df_reduced.shape
```

[70]:
```
(3985, 61)
```

[71]:
```
df_reduced.columns
```

[71]:
```
Index(['patient_barcode', 'tissue_source_site', 'patient_id',
       'bcr_patient_uuid', 'informed_consent_verified', 'icd_o_3_site',
       'icd_o_3_histology', 'icd_10',
       'tissue_prospective_collection_indicator',
       'tissue_retrospective_collection_indicator', 'days_to_birth',
       'country_of_birth', 'gender', 'height', 'weight',
       'country_of_procurement', 'state_province_of_procurement',
       'city_of_procurement', 'race_list', 'other_dx',
       'history_of_neoadjuvant_treatment', 'person_neoplasm_cancer_status',
       'vital_status', 'days_to_last_followup', 'tobacco_smoking_history',
       'number_pack_years_smoked', 'alcohol_history_documented',
       'frequency_of_alcohol_consumption',
       'amount_of_alcohol_consumption_per_day', 'reflux_history',
       'initial_diagnosis_by', 'barretts_esophagus',
       'history_of_esophageal_cancer', 'has_new_tumor_events_information',
       'day_of_form_completion', 'month_of_form_completion',
       'year_of_form_completion', 'has_follow_ups_information',
       'has_drugs_information', 'has_radiations_information', 'project',
       'stage_event_system_version', 'stage_event_pathologic_stage',
       'stage_event_tnm_categories', 'primary_pathology_tumor_tissue_site',
       'primary_pathology_esophageal_tumor_cental_location',
       'primary_pathology_esophageal_tumor_involvement_sites',
       'primary_pathology_histological_type',
       'primary_pathology_columnar_metaplasia_present',
       'primary_pathology_neoplasm_histologic_grade',
       'primary_pathology_days_to_initial_pathologic_diagnosis',
       'primary_pathology_age_at_initial_pathologic_diagnosis',
```

```
       'primary_pathology_year_of_initial_pathologic_diagnosis',
       'primary_pathology_initial_pathologic_diagnosis_method',
       'primary_pathology_lymph_node_metastasis_radiographic_evidence',
       'primary_pathology_primary_lymph_node_presentation_assessment',
       'primary_pathology_lymph_node_examined_count',
       'primary_pathology_number_of_lymphnodes_positive_by_he',
       'primary_pathology_residual_tumor',
       'primary_pathology_radiation_therapy',
       'primary_pathology_postoperative_rx_tx'],
      dtype='object')
```

[72]:
```
non_categorical_features
```

[72]:
```
['patient_barcode',
 'tissue_source_site',
 'patient_id',
 'bcr_patient_uuid',
 'country_of_procurement',
 'state_province_of_procurement',
 'city_of_procurement',
 'stage_event_pathologic_stage',
 'stage_event_tnm_categories']
```

[73]:
```
categorical_features =␣
 ↪['informed_consent_verified','icd_o_3_site','icd_o_3_histology','icd_10',
                          ␣
 ↪'tissue_prospective_collection_indicator','tissue_retrospective_collection_indicator',
                          ␣
 ↪'country_of_birth','gender','race_list','other_dx','history_of_neoadjuvant_treatment',
                          ␣
 ↪'vital_status','alcohol_history_documented','reflux_history','initial_diagnosis_by',
                          ␣
 ↪'barretts_esophagus','history_of_esophageal_cancer','has_new_tumor_events_information',
                          ␣
 ↪'has_follow_ups_information','has_drugs_information','has_radiations_information',
                          ␣
 ↪'project','stage_event_system_version','primary_pathology_tumor_tissue_site',
                           'primary_pathology_esophageal_tumor_cental_location',
                           'primary_pathology_esophageal_tumor_involvement_sites',
                          ␣
 ↪'primary_pathology_histological_type','primary_pathology_columnar_metaplasia_present',
                          ␣
 ↪'primary_pathology_neoplasm_histologic_grade','primary_pathology_initial_pathologic_diagnos
                          ␣
 ↪'primary_pathology_primary_lymph_node_presentation_assessment',
```

```
                        ␣
 ↪'primary_pathology_lymph_node_metastasis_radiographic_evidence',
                        ␣
 ↪'primary_pathology_residual_tumor','primary_pathology_radiation_therapy',
                         'primary_pathology_postoperative_rx_tx',
                         'patient_barcode', 'tissue_source_site', 'patient_id',␣
 ↪'bcr_patient_uuid', 'country_of_procurement',
                         'state_province_of_procurement', 'city_of_procurement',
                         'stage_event_pathologic_stage',␣
 ↪'stage_event_tnm_categories',
                         ]

df_reduced = pd.get_dummies(df_reduced, columns=categorical_features,␣
 ↪drop_first=True)
```

[74]:
```
df_reduced
```

[74]:
```
      days_to_birth      height     weight person_neoplasm_cancer_status  \
0             -24487  183.000000   95.00000                    WITH TUMOR
1             -24328  178.000000   74.00000                    WITH TUMOR
2             -16197  183.000000   91.00000                    WITH TUMOR
3             -25097  188.000000  100.00000                    WITH TUMOR
4             -21180  189.000000   70.00000                    WITH TUMOR
…              …         …        …                            …
3980          -19505  169.000000   67.00000                Not Available
3981          -20861  165.000000   65.00000                   TUMOR FREE
3982          -19438  162.000000   57.00000                   TUMOR FREE
3983          -20850  161.000000   69.00000                   TUMOR FREE
3984          -23820  172.128518   75.62256                Not Available

      days_to_last_followup  tobacco_smoking_history  \
0                306.201937                      1.0
1                306.201937                      1.0
2                306.201937                      1.0
3                306.201937                      1.0
4                306.201937                      1.0
…                     …                        …
3980              20.000000                      1.0
3981               0.000000                      1.0
3982              15.000000                      1.0
3983              18.000000                      4.0
3984             551.000000                      1.0

      number_pack_years_smoked  frequency_of_alcohol_consumption  \
0                    35.392577                               7.0
1                    35.392577                               7.0
2                    35.392577                               7.0
```

```
3                    35.392577                                7.0
4                    35.392577                                7.0
…                        …                                    …
3980                 35.392577                                7.0
3981                 35.392577                                7.0
3982                 35.392577                                7.0
3983                  1.000000                                7.0
3984                 35.392577                                7.0

      amount_of_alcohol_consumption_per_day  day_of_form_completion  …  \
0                                  1.749051                      25  …
1                                  1.749051                      25  …
2                                  1.749051                      25  …
3                                  1.749051                      25  …
4                                  1.749051                      25  …
…                                     …                       …  …
3980                               1.749051                      20  …
3981                               1.749051                      20  …
3982                               1.749051                      13  …
3983                               1.749051                      13  …
3984                               1.749051                      30  …

      stage_event_tnm_categories_T3NXM0T3N0M0  \
0                                       False
1                                       False
2                                       False
3                                       False
4                                       False
…                                       …
3980                                    False
3981                                    False
3982                                    False
3983                                    False
3984                                    False

      stage_event_tnm_categories_T3NXM0T3N2M0  \
0                                       False
1                                       False
2                                       False
3                                       False
4                                       False
…                                       …
3980                                    False
3981                                    False
3982                                    False
3983                                    False
3984                                    False
```

```
      stage_event_tnm_categories_T3NXM0T3N3M0  \
0                                       False
1                                       False
2                                       False
3                                       False
4                                       False
…                                         …
3980                                    False
3981                                    False
3982                                    False
3983                                    False
3984                                    False

      stage_event_tnm_categories_T3NXM0T3N3M1  \
0                                       False
1                                       False
2                                       False
3                                       False
4                                       False
…                                         …
3980                                    False
3981                                    False
3982                                    False
3983                                    False
3984                                    False

      stage_event_tnm_categories_T4N0M0  \
0                                 False
1                                 False
2                                 False
3                                 False
4                                 False
…                                   …
3980                              False
3981                              False
3982                              False
3983                              False
3984                              False

      stage_event_tnm_categories_T4N0M0T4N0M0  \
0                                       False
1                                       False
2                                       False
3                                       False
4                                       False
…                                         …
```

```
3980                                    False
3981                                    False
3982                                    False
3983                                    False
3984                                    False

      stage_event_tnm_categories_T4N1M0T4N1M0  \
0                                       False
1                                       False
2                                       False
3                                       False
4                                       False
…                                       …
3980                                    False
3981                                    False
3982                                    False
3983                                    False
3984                                    False

      stage_event_tnm_categories_T4NXM1  stage_event_tnm_categories_T4aNXMX  \
0                                 False                                False
1                                 False                                False
2                                 False                                False
3                                 False                                False
4                                 False                                False
…                                 …                                    …
3980                              False                                False
3981                              False                                False
3982                              False                                False
3983                              False                                False
3984                              False                                False

      stage_event_tnm_categories_TXNXM1
0                                 False
1                                 False
2                                 False
3                                 False
4                                 False
…                                 …
3980                              False
3981                              False
3982                              False
3983                              False
3984                              False

[3985 rows x 4589 columns]
```

[75]:
```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()

target_feature ='person_neoplasm_cancer_status'

df_reduced[target_feature] = label_encoder.
 ↪fit_transform(df_reduced[target_feature])

df1=df_reduced.copy()

df1['person_neoplasm_cancer_status'].value_counts()
```

[75]:
```
person_neoplasm_cancer_status
1    2235
2    1415
0     335
Name: count, dtype: int64
```

[76]:
```
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import RandomOverSampler

X=df1.drop(columns=['person_neoplasm_cancer_status'])
y=df1['person_neoplasm_cancer_status']

categorical_columns = X.select_dtypes(include=['object']).columns
print("Categorical columns:", categorical_columns)
```

```
Categorical columns: Index([], dtype='object')
```

[77]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)

ros = RandomOverSampler(random_state=42)
X_train_balanced, y_train_balanced = ros.fit_resample(X_train, y_train)

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix,␣
 ↪accuracy_score

logreg = LogisticRegression(random_state=42, max_iter=1000)

logreg.fit(X_train_balanced, y_train_balanced)

LogisticRegression(max_iter=1000, random_state=42)

y_pred = logreg.predict(X_test)
```

```
print("Confusion matrix:")
print(confusion_matrix(y_test, y_pred))

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

print("\nAccuracy Score:")
print(accuracy_score(y_test, y_pred))
```

```
Confusion matrix:
[[ 57   5   3]
 [ 86 319  60]
 [ 40  65 162]]

Classification Report:
              precision    recall  f1-score   support

           0       0.31      0.88      0.46        65
           1       0.82      0.69      0.75       465
           2       0.72      0.61      0.66       267

    accuracy                           0.68       797
   macro avg       0.62      0.72      0.62       797
weighted avg       0.75      0.68      0.69       797


Accuracy Score:
0.6750313676286073
```

```
c:\Users\DELL\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:469:
ConvergenceWarning:

lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
```