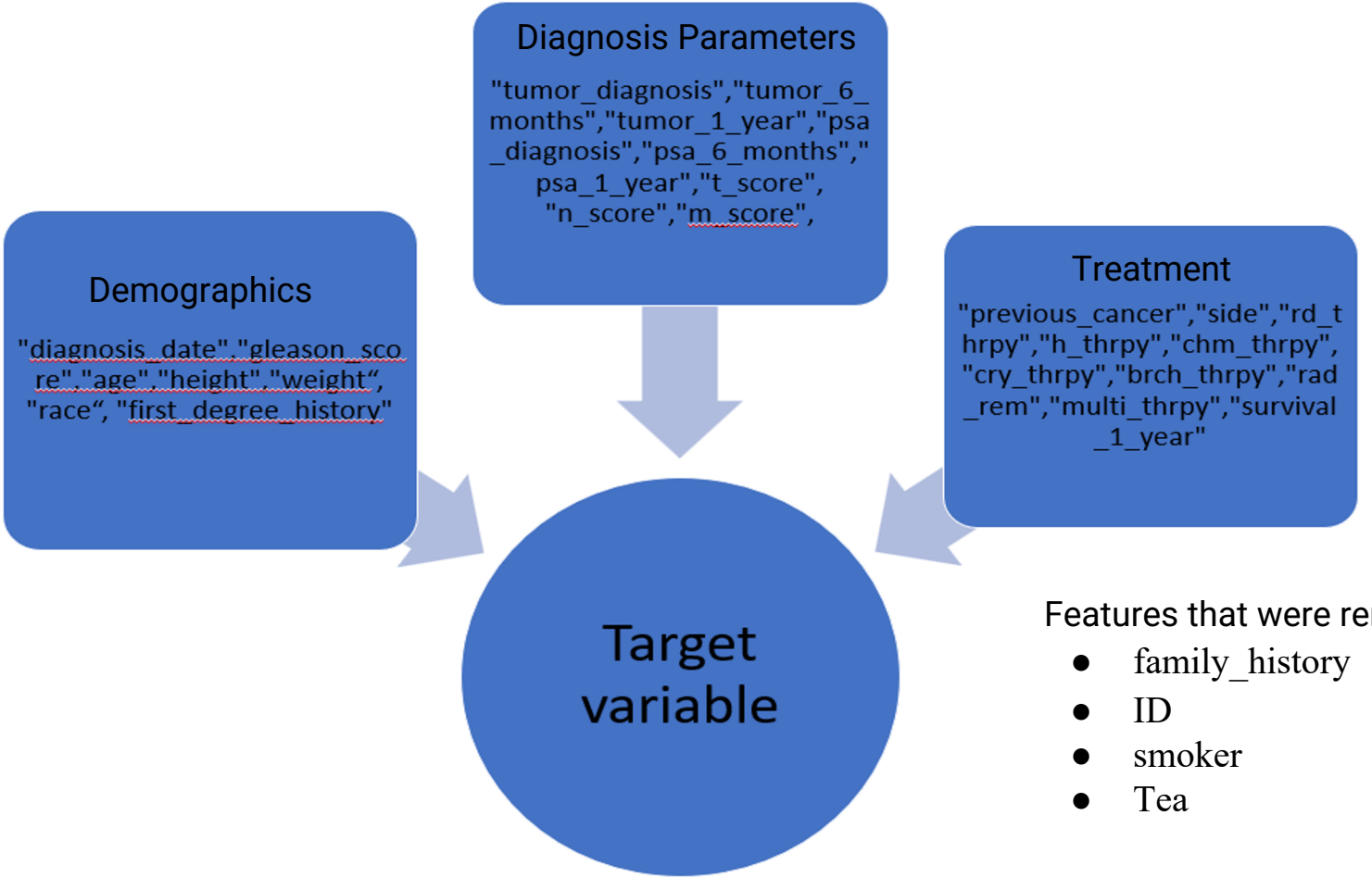
Prostate Cancer Detection

Introduction

- Prostate Cancer is the most common cancer among men, however it may often be treated successfully.
- Age, ethnicity, geography, family history, diet, obesity, smoking, and other environmental factors all the contributing risks of developing Prostate Cancer.
- In the Prostate Cancer data set (n = 15385), we develop a model to predict whether or not an individual will survive 7 years beyond initial diagnosis.
- Binary target variable - Survive 7 years (0/1)



Features that were removed:

- family_history
- ID
- smoker
- Tea

Exploratory Data Analysis

- Data shape (15385, 33)

- Categorical variables

  't_score','n_score','m_score','stage','race','first_degree_history','previous_cancer','side','rd_thrpy','h_thrpy','chm_thrpy','cry_thrpy','brch_thrpy','rad_rem','multi_thrpy','survival_1_year','survival_7_years'

- Numerical variables

  'gleason_score','age','height','weight','tumor_diagnosis','tumor_6_months','tumor_1_year','psa_diagnosis','psa_6_months','psa_1_year'

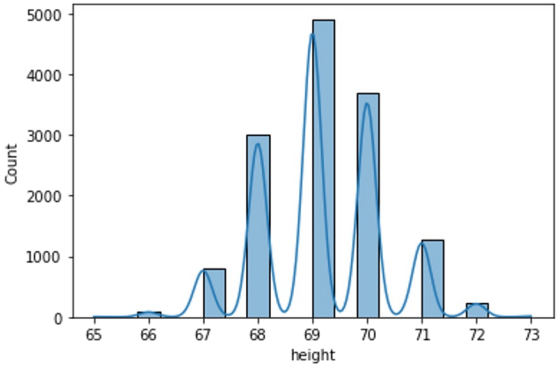
- Check for null values.

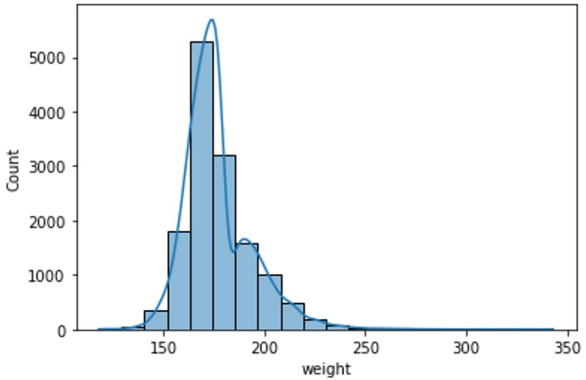
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15385 entries, 0 to 15384
Data columns (total 33 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   id                     15385 non-null  int64
 1   diagnosis_date         15385 non-null  object
 2   gleason_score          15065 non-null  float64
 3   t_score                15385 non-null  object
 4   n_score                15385 non-null  object
 5   m_score                15385 non-null  object
 6   stage                  15385 non-null  object
 7   age                    14637 non-null  float64
 8   race                   15220 non-null  float64
 9   height                 14021 non-null  float64
 10  weight                 14068 non-null  float64
 11  family_history         13799 non-null  float64
 12  first_degree_history   13799 non-null  float64
 13  previous_cancer        13799 non-null  float64
 14  smoker                 13799 non-null  float64
 15  side                   15385 non-null  object
 16  tumor_diagnosis        15082 non-null  float64
 17  tumor_6_months         5322 non-null   float64
 18  tumor_1_year           13262 non-null  float64
 19  psa_diagnosis          13987 non-null  float64
 20  psa_6_months           5882 non-null   float64
 21  psa_1_year             12868 non-null  float64
 22  tea                    13799 non-null  float64
 23  symptoms               14975 non-null  object
 24  rd_thrpy               15385 non-null  int64
 25  h_thrpy                15385 non-null  int64
 26  chm_thrpy              15385 non-null  int64
 27  cry_thrpy              15385 non-null  int64
 28  brch_thrpy             15385 non-null  int64
 29  rad_rem                15385 non-null  int64
 30  multi_thrpy            15385 non-null  int64
 31  survival_1_year        15385 non-null  int64
 32  survival_7_years       15385 non-null  int64
dtypes: float64(16), int64(10), object(7)
```

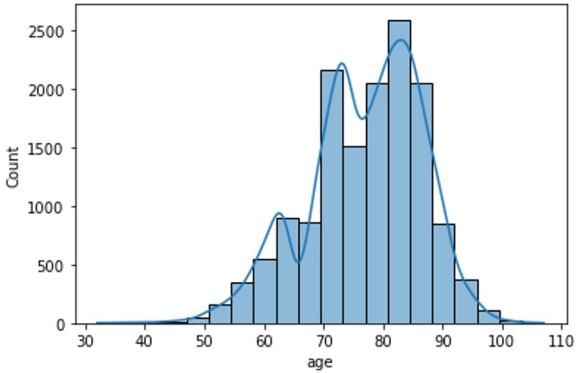
```
dfa.isnull().sum()
```

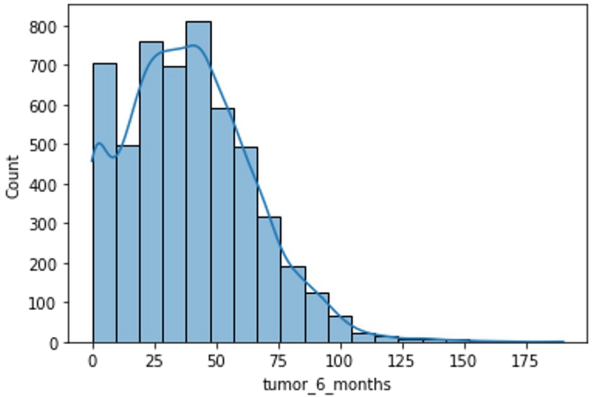
```
id                          0
diagnosis_date              0
gleason_score             320
t_score                     0
n_score                     0
m_score                     0
stage                       0
age                       748
race                      165
height                   1364
weight                   1317
family_history           1586
first_degree_history     1586
previous_cancer          1586
smoker                   1586
side                        0
tumor_diagnosis           303
tumor_6_months          10063
tumor_1_year             2123
psa_diagnosis            1398
psa_6_months             9503
psa_1_year               2517
tea                      1586
symptoms                  410
rd_thrpy                    0
h_thrpy                     0
chm_thrpy                   0
cry_thrpy                   0
brch_thrpy                  0
rad_rem                     0
multi_thrpy                 0
survival_1_year             0
survival_7_years            0
dtype: int64
```

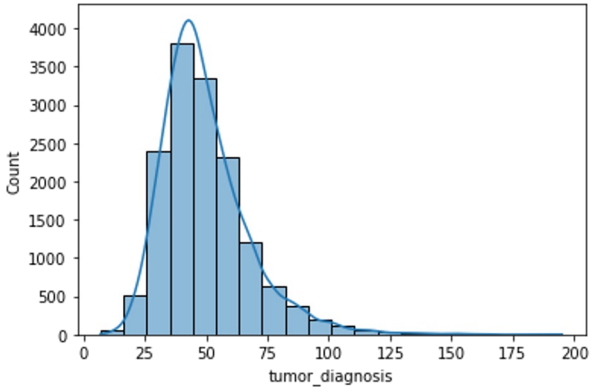
Distributions of Features

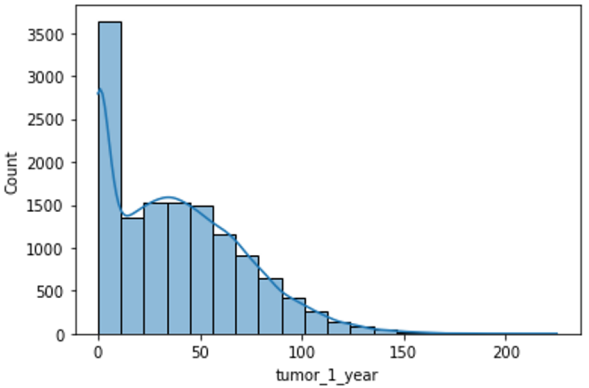




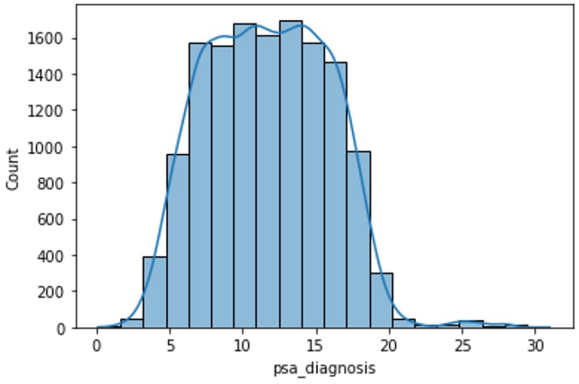


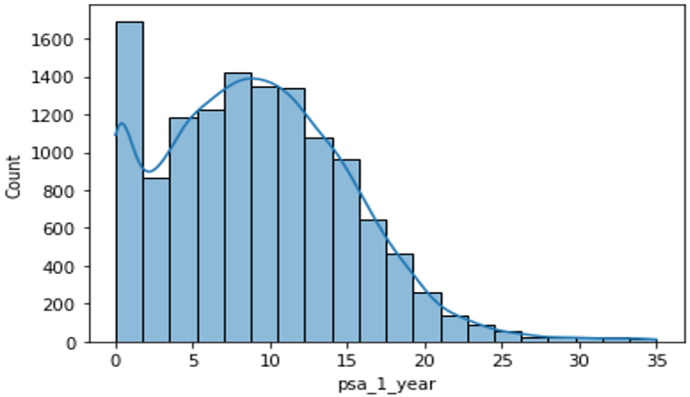


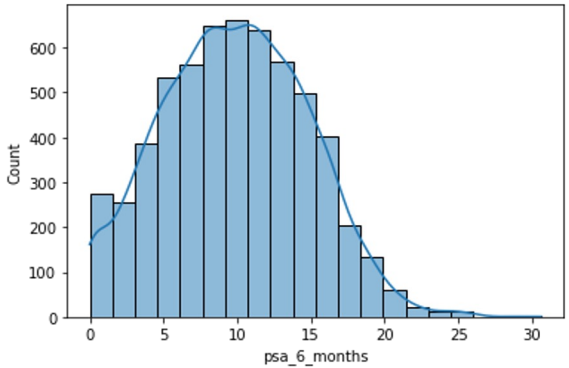


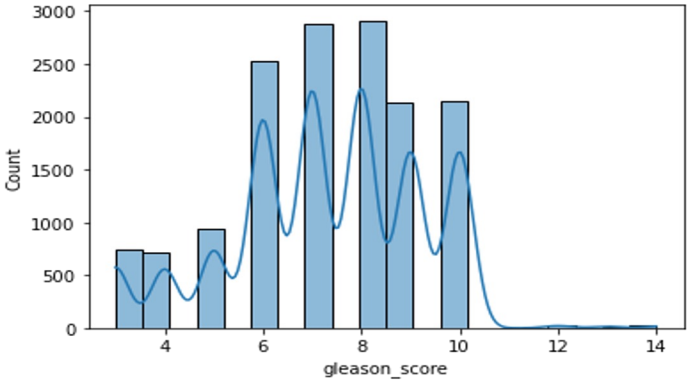


Distributions of Features

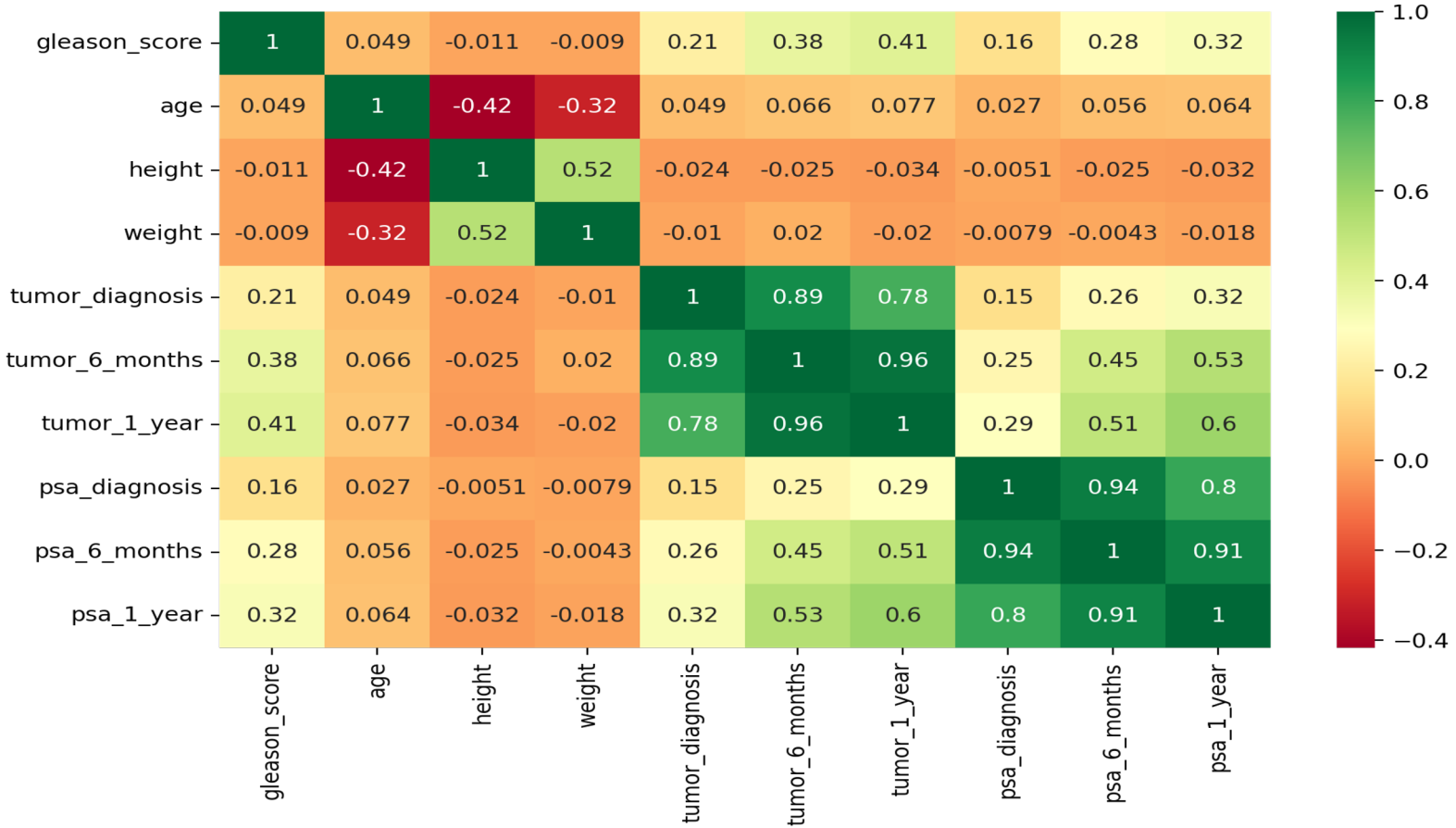




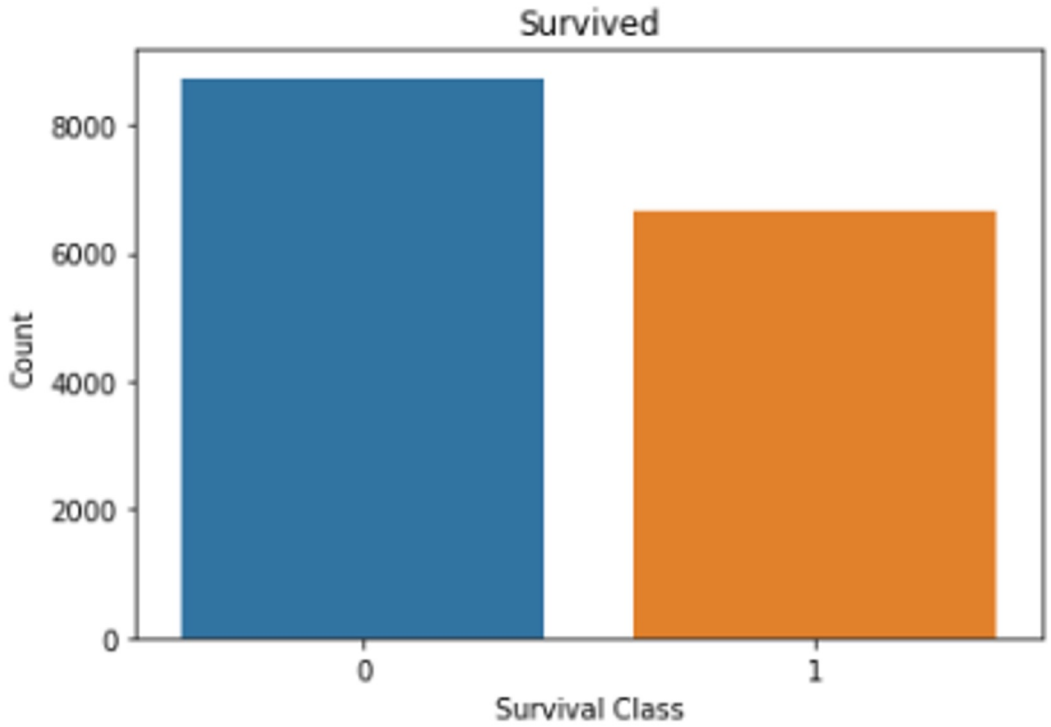




Correlation heatmap of numerical variables

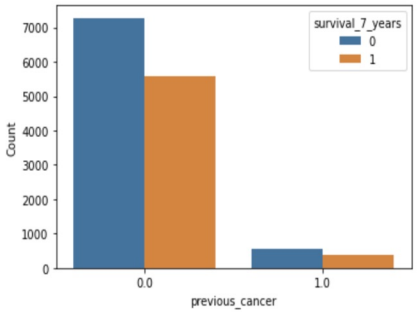


Survival Rate

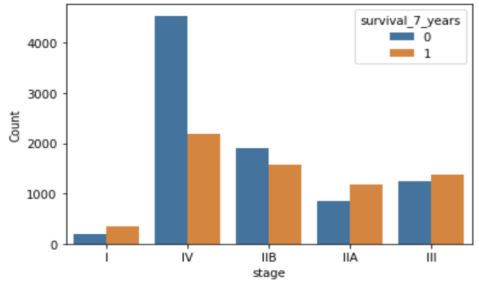



```
Percent of Survived:  43.23 %
Percent of who did not survive:  56.77 %
```

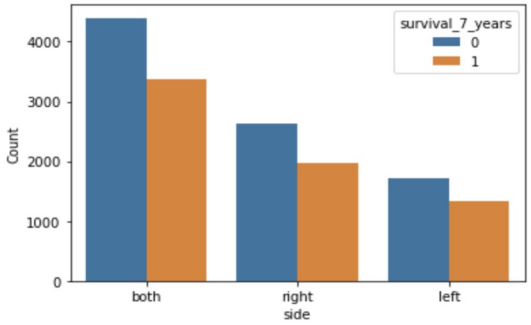
Chi-squared Test




```
observed data:
 survival_7_years     0     1
previous_cancer
0.0                7273  5577
1.0                 575   374
expected data:
 [[7308.26871512 5541.73128488]
 [ 539.73128488  409.26871512]]
chi-squared value: 5.7386 for 1 dof; p-value = 0.0166
```




```
observed data:
 survival_7_years     0     1
stage
I                  185   348
IIA                861  1184
IIB               1896  1567
III               1255  1376
IV                4537  2176
expected data:
 [[ 302.58186545  230.41813455]
 [1160.93792655  884.06207345]
 [1965.93058174 1497.06941826]
 [1493.60766981 1137.39233019]
 [3810.94195645 2902.05804355]]
chi-squared value: 698.8528 for 4 dof; p-value = 0.0000
```



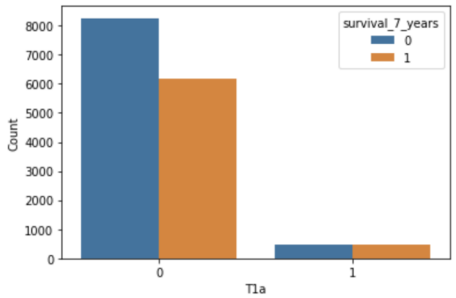

```
observed data:
 survival_7_years     0     1
side
both              4397  3364
left              1705  1323
right             2632  1964
expected data:
 [[4405.88716282 3355.11283718]
 [1718.98290543 1309.01709457]
 [2609.12993175 1986.87006825]]
chi-squared value: 0.7683 for 2 dof; p-value = 0.6810
```

- Significant variables: t_score, m_score, n_score, stage, race, first_degree_hist, previous_cancer, rd_thrpy, h_thrpy, chm_thrpy, cry_thrpy, brch_thrpy, rad_rem, multi_thrpy, survival_1_year
- Removed Insignificant variables like side.

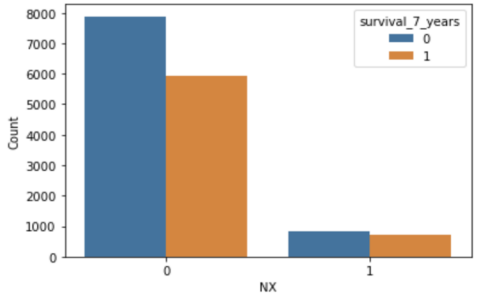
Feature Engineering

- Apply label encoder to each column with categorical data
- Created dummy variables for the following categorical variables:
    - t_score (t1a, t1b, t1c,...,t4)
    - m_score (m0, m1a, m1b, m1c)
    - n_score (n0, n1, nx)
    - stage (I, IIA, IIB, III, IV)
    - race (race_1.0, race_2.0, race_3.0, race_4.0)
    - famhis (famhis_1.0, famhis_2.0, famhis_3.0, famhis_4.0)
    - Performed chi square test to check between newly created dummy variables and target variable.
- Converted the height and weight columns to BMI.
- Seperated the diagnosis_date to month and year.
- Converted the month variable to numerical by assigning values from 1 - 12.

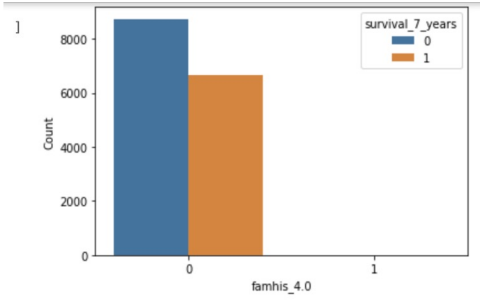
Chi-squared Test on Dummy Variables




```
observed data:
 survival_7_years     0     1
T1a
0                 8260  6188
1                  474   463
expected data:
 [[8202.06902827 6245.93097173]
 [ 531.93097173  405.06902827]]
chi-squared value: 15.5406 for 1 dof; p-value = 0.0001
```




```
observed data:
 survival_7_years     0     1
NX
0                 7885  5942
1                  849   709
expected data:
 [[7849.52993175 5977.47006825]
 [ 884.47006825  673.52993175]]
chi-squared value: 3.6612 for 1 dof; p-value = 0.0557
```




```
observed data:
 survival_7_years     0     1
famhis_4.0
0                 8731  6650
1                    3     1
expected data:
 [[8.73172922e+03 6.64927078e+03]
 [2.27078323e+00 1.72921677e+00]]
chi-squared value: 0.5418 for 1 dof; p-value = 0.4617
```

- Insignificant variables (where p-value>0.05) can be removed.
- "famhis_2.0 ","famhis_3.0 ","famhis_4.0" and “NX” were removed.
- Final dataset has 7692 rows and 43 features (dropping all the null values).

Model - Logistic Regression- 1

```
lr_acc = accuracy_score(y_train, lr.predict(X_train))
print(f"Accuracy Score is {lr_acc}")
```

```
Accuracy Score is 0.6324752990119604
```

	feature	coefficient	odds_ratio
0	gleason_score	-0.218116	0.804032
1	age	0.005241	1.005255
2	previous_cancer	0.008494	1.008530
3	tumor_diagnosis	0.126262	1.134580
4	tumor_1_year	-0.392489	0.675374
5	psa_diagnosis	-0.055345	0.946158
6	psa_1_year	0.101390	1.106708
7	rd_thrpy	-0.170927	0.842883
8	h_thrpy	0.014541	1.014648
9	chm_thrpy	0.038519	1.039271
10	cry_thrpy	-0.017726	0.982430
11	brch_thrpy	-0.103444	0.901727
12	rad_rem	-0.092469	0.911678
13	multi_thrpy	-0.045922	0.955117
14	survival_1_year	0.000000	1.000000
15	T1a	-0.043357	0.957570
16	T1b	-0.030326	0.970129
17	T1c	-0.059483	0.942251
18	T2a	-0.030217	0.970235
19	T2b	-0.028956	0.971459
20	T2c	-0.088497	0.915305
21	T3a	0.040978	1.041830
22	T3b	0.035051	1.035673
23	T3c	0.040508	1.041339
24	T4	0.089151	1.093246
25	N0	-0.021790	0.978446
26	N1	-0.393022	0.675014
27	M0	0.112599	1.119183
28	M1a	-0.046661	0.954410
29	M1b	-0.076333	0.926508
30	M1c	-0.070367	0.932052
31	I	0.062075	1.064042
32	IIA	0.097154	1.102030
33	IIB	0.079410	1.082649
34	III	0.016396	1.016532
35	IV	-0.172968	0.841164
36	race_1.0	-0.044725	0.956260
37	famhis_0.0	-0.086951	0.916722
38	famhis_1.0	-0.084847	0.918653
39	bmi	-0.037474	0.963219
40	month	0.007819	1.007850
41	year	0.008902	1.008941

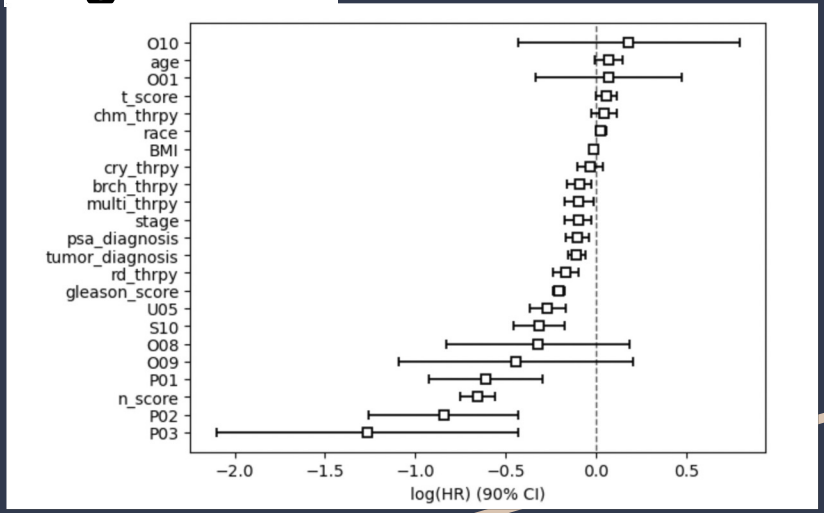
Model - Logistic Regression -2

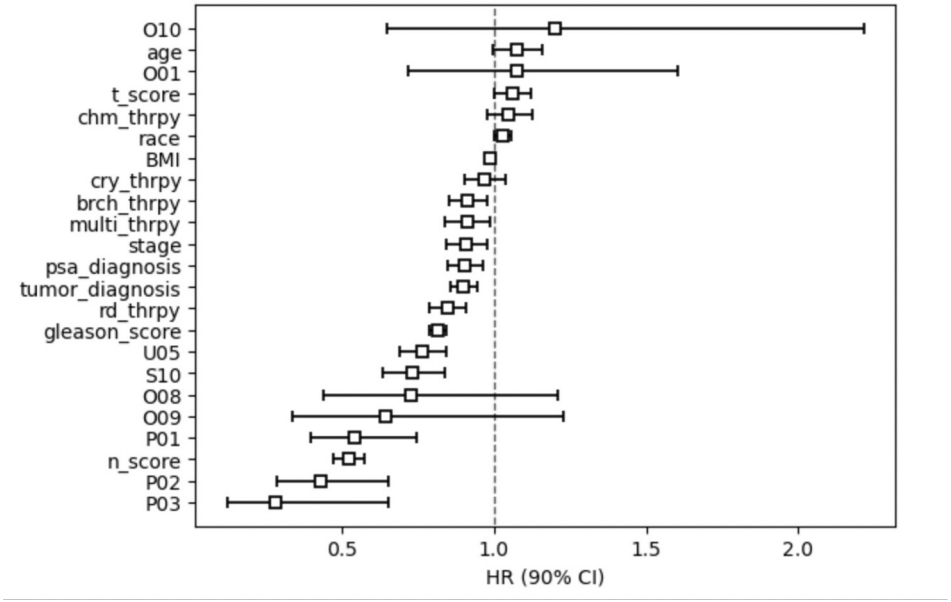
	precision	recall	f1-score	support
0	0.66	0.73	0.69	1455
1	0.58	0.49	0.53	1091
accuracy			0.63	2546
macro avg	0.62	0.61	0.61	2546
weighted avg	0.62	0.63	0.62	2546

```
[[1062  393]
 [ 553  538]]
```

	feature	coefficient	odds_ratio
0	age	0.230653	1.259422
1	gleason_score	-0.323337	0.723730
2	t_score	0.041247	1.042109
3	n_score	-1.024488	0.358980
4	stage	-0.044537	0.956440
5	race	0.048436	1.049628
6	tumor_diagnosis	-0.199989	0.818739
7	psa_diagnosis	-0.138882	0.870330
8	rd_thrpy	-0.319498	0.726514
9	h_thrpy	0.162205	1.176102
10	chm_thrpy	0.103613	1.109171
11	cry_thrpy	0.034776	1.035387
12	brch_thrpy	-0.005890	0.994127
13	multi_thrpy	-0.294818	0.744667
14	O01	-0.230629	0.794034
15	O08	-0.466727	0.627051
16	O09	-0.573860	0.563347
17	O10	-0.167400	0.845862
18	P01	-0.710263	0.491515
19	P02	-0.700025	0.496573
20	S04	0.116307	1.123341
21	P03	-0.380319	0.683643
22	S10	-0.614781	0.540759
23	U05	-0.302307	0.739111
24	BMI	0.004433	1.004443

Survival Analysis - Cox Proportional Hazards Regression





```
CoxFitter.print_summary(decimals=3)
```

model	lifelines.CoxPHFitter
duration col	'tenure'
event col	'survival_7_years'
baseline estimation	breslow
number of observations	8485
number of events observed	3650
partial log-likelihood	-31160.639
time fit was run	2023-03-07 19:34:20 UTC

Concordance	0.678
Partial AIC	62367.278
log-likelihood ratio test	780.181 on 23 df
-log2(p) of ll-ratio test	495.866

	coef	exp(coef)	se(coef)	coef lower 90%	coef upper 90%	exp(coef) lower 90%	exp(coef) upper 90%	cmp to	z	p	-log2(p)
age	0.071	1.074	0.046	-0.004	0.147	0.996	1.158	0.000	1.552	0.121	3.050
gleason_score	-0.204	0.815	0.020	-0.237	-0.172	0.789	0.842	0.000	-10.314	<0.0005	80.435
t_score	0.059	1.060	0.035	0.002	0.116	1.002	1.122	0.000	1.698	0.089	3.482
n_score	-0.654	0.520	0.060	-0.752	-0.556	0.472	0.574	0.000	-10.982	<0.0005	90.794
stage	-0.098	0.907	0.046	-0.173	-0.023	0.841	0.978	0.000	-2.143	0.032	4.962
race	0.028	1.028	0.018	-0.001	0.057	0.999	1.058	0.000	1.577	0.115	3.122
tumor_diagnosis	-0.106	0.900	0.030	-0.154	-0.057	0.857	0.945	0.000	-3.569	<0.0005	11.448
psa_diagnosis	-0.102	0.903	0.039	-0.165	-0.038	0.848	0.963	0.000	-2.627	0.009	6.860
rd_thrpy	-0.166	0.847	0.043	-0.237	-0.095	0.789	0.910	0.000	-3.847	<0.0005	13.029
chm_thrpy	0.047	1.048	0.043	-0.023	0.117	0.977	1.125	0.000	1.103	0.270	1.888
cry_thrpy	-0.033	0.968	0.042	-0.103	0.037	0.902	1.038	0.000	-0.777	0.437	1.194
brch_thrpy	-0.092	0.912	0.042	-0.161	-0.023	0.851	0.977	0.000	-2.188	0.029	5.126
multi_thrpy	-0.094	0.911	0.050	-0.176	-0.012	0.839	0.988	0.000	-1.880	0.060	4.057
O01	0.071	1.074	0.245	-0.332	0.474	0.718	1.606	0.000	0.290	0.772	0.374
O08	-0.319	0.727	0.309	-0.827	0.190	0.437	1.209	0.000	-1.031	0.303	1.724
O09	-0.443	0.642	0.395	-1.093	0.207	0.335	1.230	0.000	-1.120	0.263	1.930
O10	0.182	1.200	0.373	-0.432	0.796	0.649	2.216	0.000	0.488	0.626	0.676
P01	-0.610	0.543	0.190	-0.924	-0.297	0.397	0.743	0.000	-3.205	0.001	9.533
P02	-0.842	0.431	0.253	-1.258	-0.427	0.284	0.653	0.000	-3.334	0.001	10.191
P03	-1.265	0.282	0.508	-2.100	-0.429	0.122	0.651	0.000	-2.490	0.013	6.292
S10	-0.314	0.730	0.084	-0.453	-0.176	0.636	0.839	0.000	-3.734	<0.0005	12.371
U05	-0.269	0.764	0.061	-0.368	-0.169	0.692	0.845	0.000	-4.423	<0.0005	16.647
BMI	-0.015	0.986	0.008	-0.027	-0.002	0.973	0.998	0.000	-1.873	0.061	4.033

Insights

- **O10, Age, O01, t_score, Chemo_thrpy, BMI, cry_thrpy** and Race are all within the 90% confidence interval of influencing the SURVIVAL event.
- The p-values from the summary tell us that **n_score, gleason score, stage , tumor diagnosis ,psa ,rd_thrpy,U05,S10**are highly significant. Their p-value is less than 0.0005, implying a statistical significance at a (100 – 0.01) = 99.99% or higher confidence level. These features are highly correlated to the SURVIVAL event

- So the above mentioned variables can be the important/major predictors