Prostate Cancer: Casestudy Analysis

Harsh Shah

Problem Statement

- To determine the 7-year survival of prostate cancer patients

Dataset:

Training: 33 variables, 15385 records.

Testing: 11531 records (predict the 7-year survivor variable)

Variable Selection

- Dropped id and diagnosis_date variables
- tumor_6_months and psa_6_months were having ~70% missing values
- stage and family_history were highly collinear

```
> sapply(Dataset,function(x) sum(is.na(x)))
      gleason_score             t_score             n_score             m_score
                320                   0                   0                   0
               race  first_degree_history    previous_cancer              smoker
                165                1586                1586                1586
               side                 tea            symptoms           rd_thrpy
                  0                1586                   0                   0
          h_thrpy           chm_thrpy           cry_thrpy          brch_thrpy
                  0                   0                   0                   0
           rad_rem          multi_thrpy       survival_1_year     survival_7_years
                  0                   0                   0                   0
               age              height              weight       tumor_diagnosis
                748                1364                1317                 303
    tumor_6_months         tumor_1_year       psa_diagnosis        psa_6_months
             10063                2123                1398                9503
        psa_1_year
              2517
```

```
> Model<-glm(formula = survival_7_years ~ ., family = "binomial", data = Dataset)
> car::vif(Model)
                        GVIF Df GVIF^(1/(2*Df))
gleason_score       2.331227 10        1.043228
t_score            38.752252  9        1.225290
n_score             2.740113  2        1.286596
m_score             1.243934  3        1.037050
stage             104.537867  4        1.788172
race                1.078436  3        1.012665
family_history      7.122855  5        1.216929
first_degree_history 7.152257  4        1.278808
previous_cancer     1.013331  1        1.006643
smoker              1.030744  1        1.015256
side                1.018668  2        1.004635
tea                 1.111757 11        1.004827
rd_thrpy            1.840549  1        1.356668
h_thrpy             1.515733  1        1.231151
chm_thrpy           1.760705  1        1.326916
cry_thrpy           1.396808  1        1.181866
brch_thrpy          1.447234  1        1.203010
rad_rem             1.454517  1        1.206033
multi_thrpy         2.265893  1        1.505288
age                 1.260314  1        1.122637
height              1.567013  1        1.251804
weight              1.427362  1        1.194722
tumor_diagnosis     2.979639  1        1.726163
tumor_1_year        5.096571  1        2.257559
psa_diagnosis       3.633585  1        1.906196
psa_1_year          5.379515  1        2.319378
```

Data Imputation

- 1586 patients didn't survive for upto 1 year after diagnosis
  1. For all these records, if either of psa_6_months or psa_1 year values was missing; it was imputed using psa_diagnosis value
  2. For all these records, if either of tumor_6_months or tumor_1 year values was missing; it was imputed using tumor_diagnosis value
- 7282 patient records for psa_6_months were computed using psa_diagnosis and psa_1 year
- 8486 patient records for tumor_6_months were computed using tumor_diagnosis and tumor_1 year

Handling Missing Values

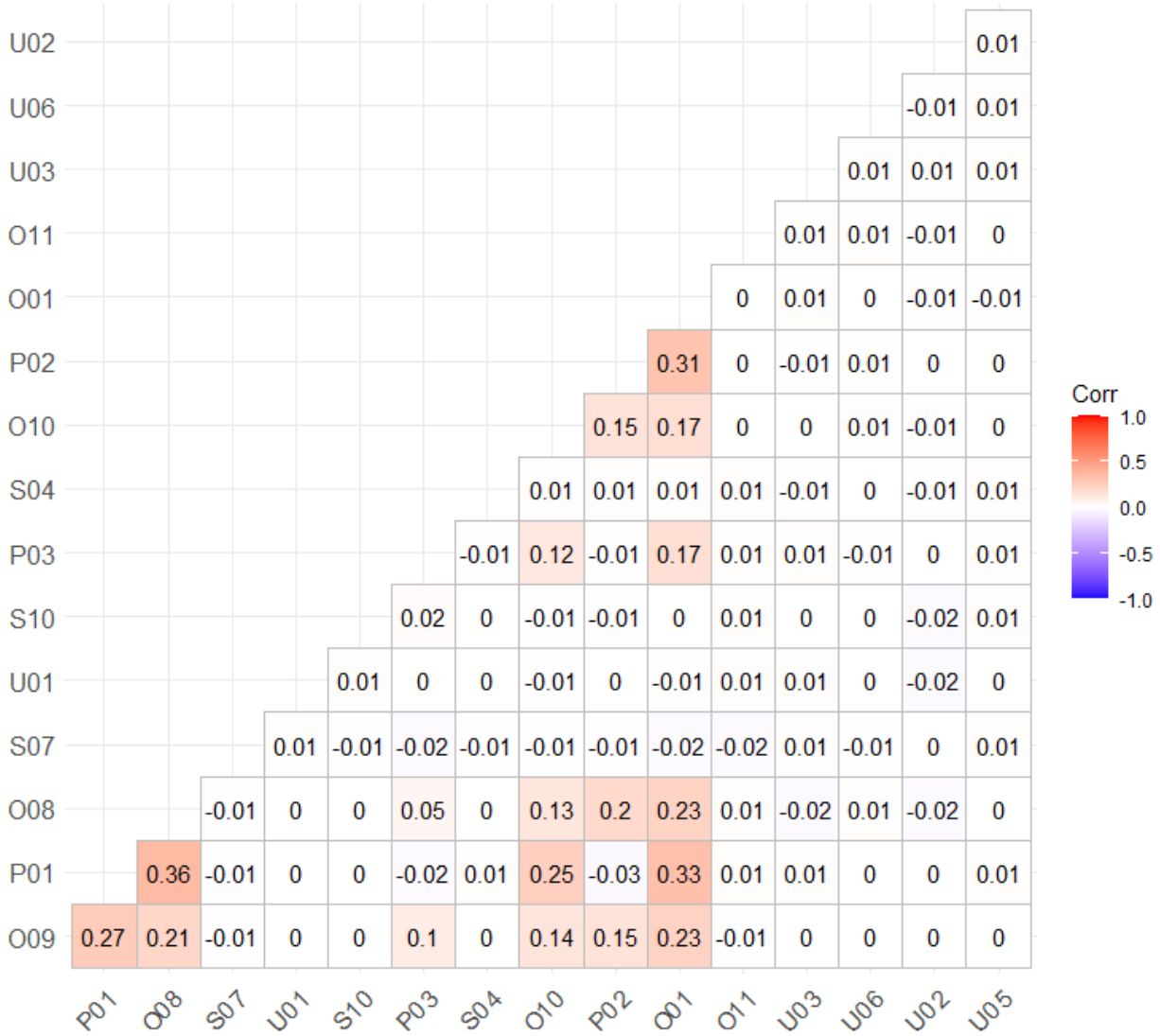
For all the demographic variables – age, race, height, weight the missing values were imputed using predictive mean matching.

Variable	Missing Values
gleason_score	320
first_degree_history	1550
tumor_1_year, tumor_6_months	427, 123
psa_diagnosis	1174
tumor_diagnosis	175
psa_1_year	691
symptoms	282

```
> sapply(Dataset,function(x) sum(is.na(x)))
      gleason_score             t_score             n_score             m_score               stage
                  0                   0                   0                   0                   0
                age                race              height              weight       family_history
                523                 112                 954                 953                   0
first_degree_history      previous_cancer              smoker                side     tumor_diagnosis
                  0                   0                   0                   0                   0
     tumor_6_months        tumor_1_year       psa_diagnosis       psa_6_months        psa_1_year
                  0                   0                   0                   0                   0
                tea            symptoms            rd_thrpy            h_thrpy           chm_thrpy
                  0                   0                   0                   0                   0
        cry_thrpy           brch_thrpy            rad_rem          multi_thrpy       Total_Therapy
                  0                   0                   0                   0                   0
    survival_1_year     survival_7_years
                  0                   0
```

Symbols Variable

- A list of codes indicating the presence of various symptoms
- Missing values were removed
- Converted the multi-valued variable into 16 different binary variables



Variable Conversion

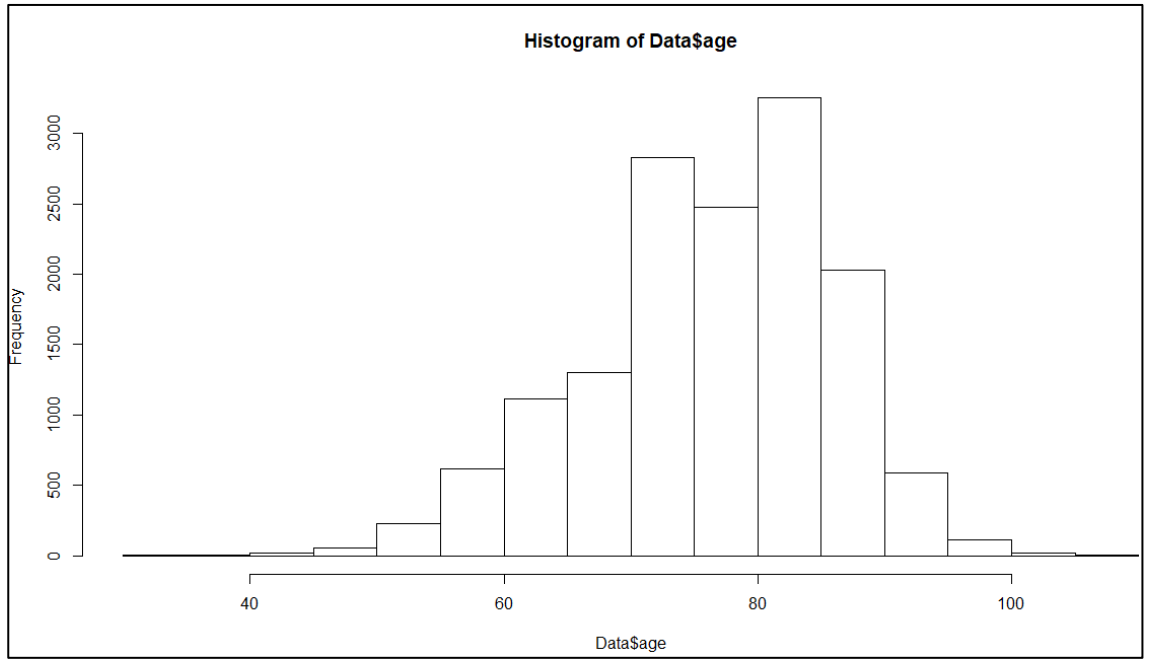
- **Age:**

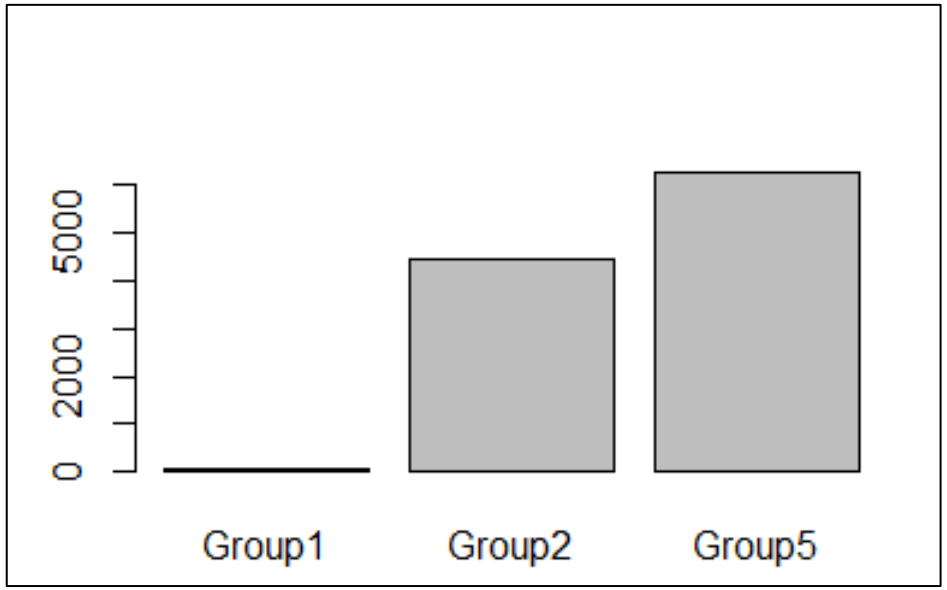
Group1 – Less than or equal to 50

Group2 –50 to 75

Group5 – Greater than 75

A new variable age_grp was created with 3 groups as shown and age was dropped.





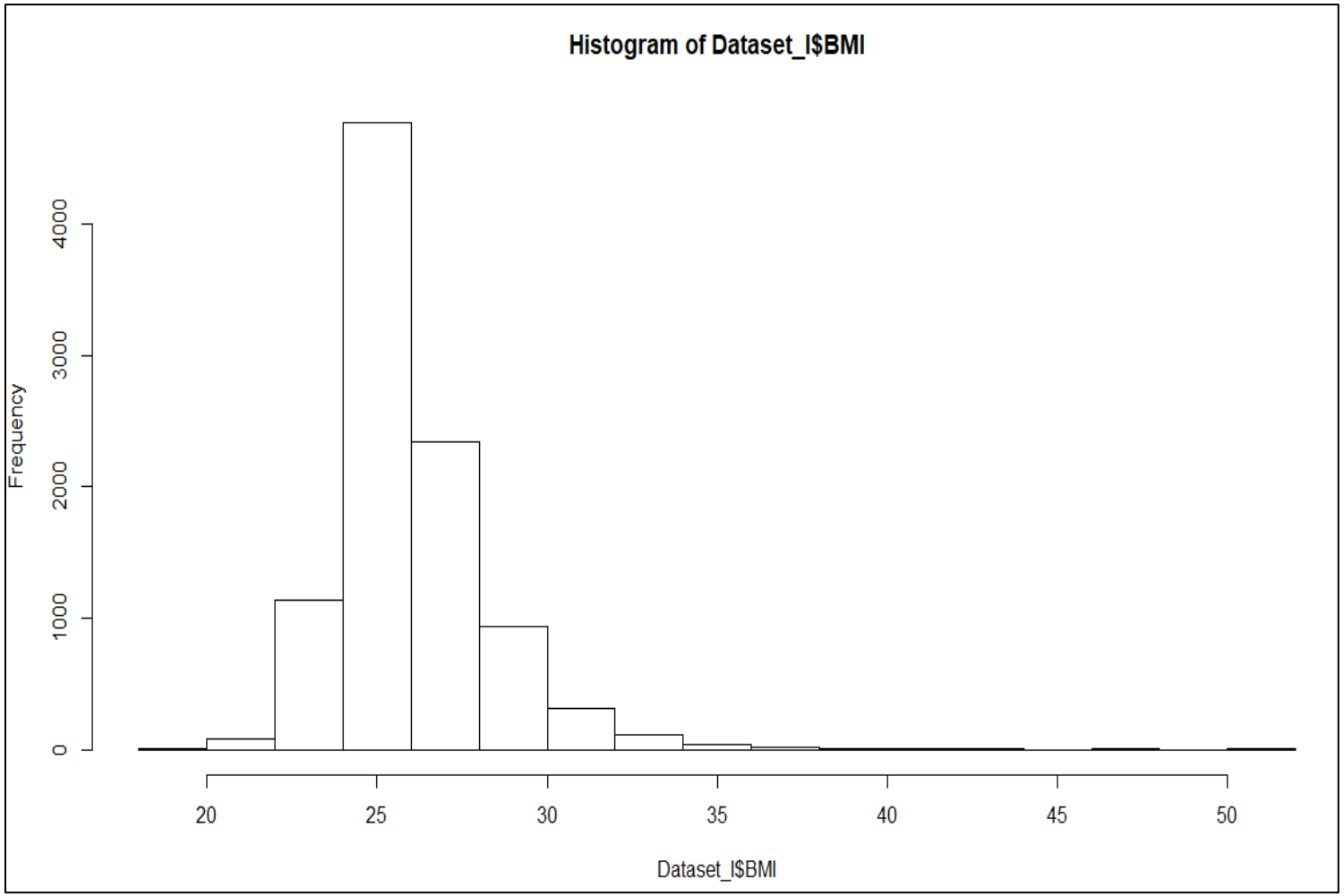
Variable Conversion

- **BMI:**

BMI was calculated using weight and height

BMI = weight/(height)^2

The variables height and weight were dropped



Variable Conversion

- **Gleason_score:**

Low – 3, 4, 5, 6

Medium – 7, 8, 9

High – 10, 11, 12, 13, 14

A new variable gleason_scr was created with 3 groups as shown and gleason_score was dropped.

7-YEAR SURVIVAL	LOW	MEDIUM	HIGH
0	1708	3381	1130
1	1809	2288	450

Variable Conversion

- **PSA growth rate:**

For 1 year, the PSA growth rate was calculated

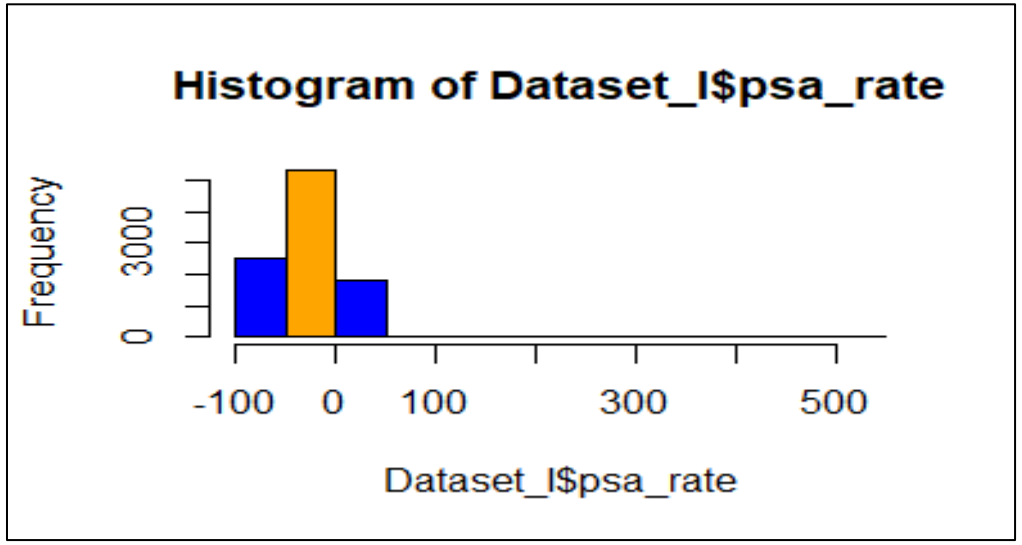
Negative values/ left skew indicate that more patients have overall decrease

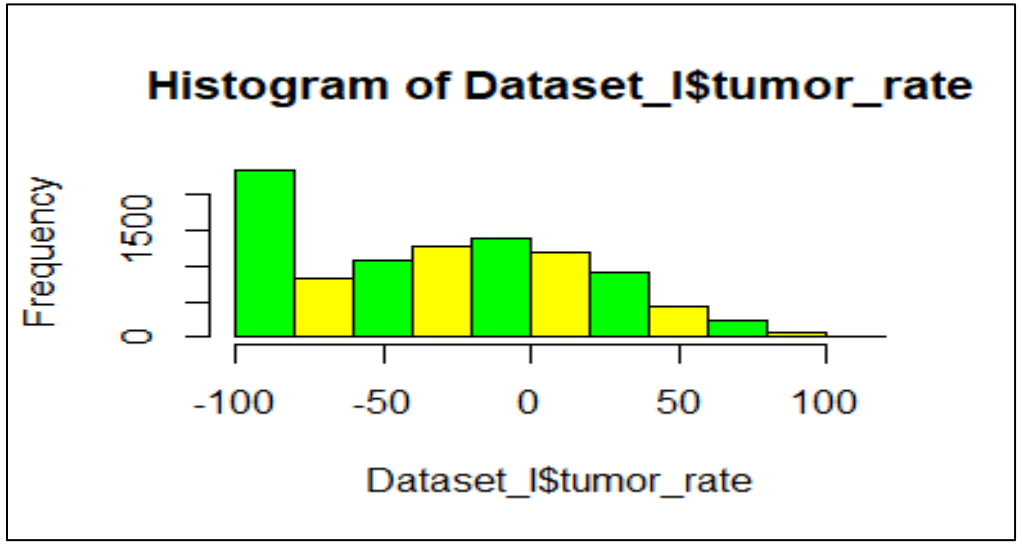
- **Tumor growth rate:**

For 1 year, the tumor growth rate was calculated

Negative values/ left skew indicate that more patients have overall decrease

psa_1_year, psa_diagnosis, tumor_diagnosis, tumor_1_year were dropped





Bi-variate Analysis (Numeric)

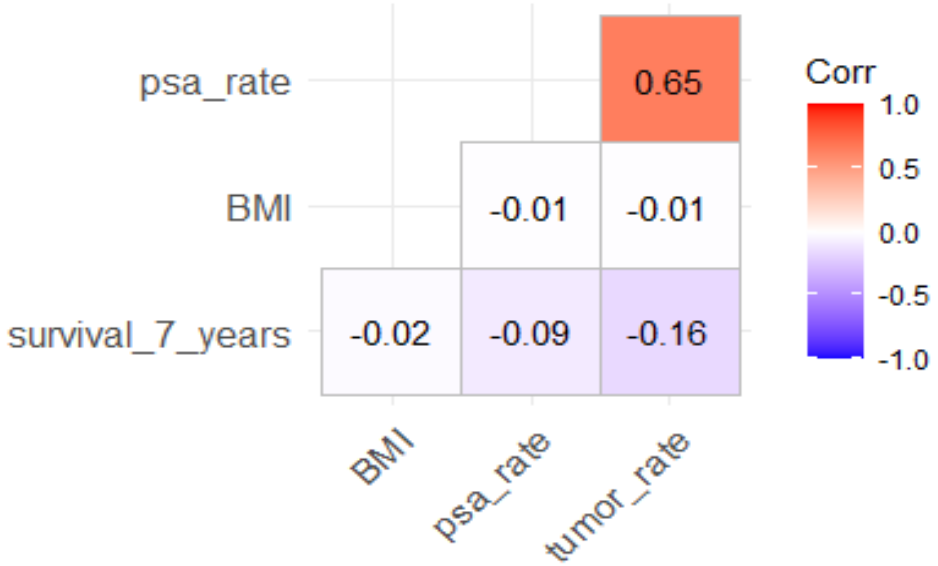
- Psr_rate
- BMI
- Tumor_rate
- Survival_7_years

Psa_rate and tumor_rate are co-related.

T-test:

H0- There is no significant different across the two groups

H1- There is significant difference across the two groups



Variables	DV Group	Mean	P-value	Significant
Psr_rate	0	-24	<0.01	Yes
	1	-32		
BMI	0	28	<0.01	Yes
	1	24		
Tumor_rate	0	-24	<0.01	Yes
	1	-40		

Bi-variate Analysis (Categorical)

Chi-squared test:

H0- There is no significant different across the two groups

H1- There is significant difference across the two groups

Variable	Variable 0	Variable 1	%1's	P-value	Significant
smoker	4828	250	5.18	0.037	No
	4404	276	6.28		
rd_thrpy	2130	2948	138.4	<0.01	Yes
	2498	2182	87.34		
chm_thrpy	1596	3482	218.17	<0.01	Yes
	1797	2883	160.43		
multi_thrpy	1020	4058	397.8	<0.01	Yes
	1185	3495	394.9		
previous_cancer	4711	367	7.79	0.17	No
	4375	305	6.97		
P02	4909	169	3.44	<0.01	Yes
	4640	40	0.86		

Survival Percentage for 1_year and 7-year wrt Age, Cancer stage and Therapies used

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	1 year		1			2			3			4			5			6	
2	Survival %	rd_thrpy			h_thrpy			chm_thrpy			cry_thrpy			brch_thrpy			rad_rem		
3	Multi_thr	100% (2,3)	98% (3)	98% (3)	100% (1,3)	99% (1,3)	99% (1,3)	82% (1)	82% (1)	83% (1)	90% (3)	87% (3)	85% (3)	71%	84% (3)	85%	83%	85%	86%
4	Age_Grp	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+
5	Stage I	NA	93%	97%	NA	95%	93%	100%	94%	96%	NA	92%	97%	100%	95%	93%	NA	96%	93%
6	Stage II	100%	90%	93%	100%	91%	93%	100%	93%	93%	89%	90%	93%	100%	94%	93%	86%	90%	92%
7	Stage III	100%	90%	91%	100%	91%	92%	100%	92%	94%	100%	88%	88%	100%	90%	93%	NA	89%	93%
8	Stage IV	87%	84%	84%	86%	86%	86%	87%	87%	86%	100%	87%	86%	80%	86%	86%	NA	72%	75%
9																			
10																			
11	7 years																		
12	Survival %	rd_thrpy			h_thrpy			chm_thrpy			cry_thrpy			brch_thrpy			rad_rem		
13	Multi_thr	100% (2)	99% (2,3)	99% (2,3)	100% (3)	99% (1,3)	99% (1,3)	75%	78%	81%	86% (3)	85% (3)	84%	60%	82%	83%	100%(3)	84%	85%
14	Age_Grp	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+	<50	50-75	75+
15	Stage I	NA	73%	64%	NA	63%	64%	100%	52%	70%	NA	57%	74%	100%	69%	62%	NA	52%	59%
16	Stage II	25%	49%	48%	25%	47%	48%	67%	50%	48%	56%	54%	50%	39%	53%	51%	14%	44%	46%
17	Stage III	14%	50%	50%	17%	51%	50%	29%	57%	59%	100%	49%	48%	100%	48%	53%	NA	44%	48%
18	Stage IV	27%	28%	24%	29%	37%	30%	39%	35%	31%	100%	37%	31%	20%	32%	28%	NA	22%	16%
19																			
20	NA - No record found																		
21	The records for age group less than 50 were not sufficient																		
22	All the patients with age less than 50 and in Stage IV were given chm_thrpy while non of them were given rad_rem																		

- For 7 years, the survival probability for a patient aged 75+ and Stage III cancer stage is 59% with chemotherapy(the highest across all different therapies). Having said that, 81% of those patients given chemotherapy underwent multiple therapies
- For 1 year, the survival probability for a patient aged between 50-75 and Stage III cancer stage is 92% with chemotherpy(the highest across all different therapies). Having said that, 82% of those patients given chemotherapy underwent multiple therapies especially external beam radiotherapy

Model : Logistic Regression

- Dataset:

19 variables, 10766 records

Train: 80%, Test: 20%

(DV proportion was maintained)

- Stepwise Regression technique and multiple regression models with different set of variables was used to determine the most significant variables

```
> Model<-glm(formula = survival_7_years ~ survival_1_year + n_score +
+              rd_thrpy + gleason_score + m_score + U05 + S10 + rad_rem +
+              brch_thrpy + tumor_1_year + stage + smoker + cry_thrpy +
+              Total_Therapy + age_grp + age + tumor_diagnosis + psa_rate +
+              psa6_rate, family = "binomial", data = TrainData)
> confusionMatrix(table(pre,TestData$survival_7_years), positive = "1")
Confusion Matrix and Statistics


pre   0   1
  0 804 252
  1 439 657

               Accuracy : 0.6789
                 95% CI : (0.6587, 0.6986)
    No Information Rate : 0.5776
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3597

 Mcnemar's Test P-Value : 1.486e-12

            Sensitivity : 0.7228
            Specificity : 0.6468
         Pos Pred Value : 0.5995
         Neg Pred Value : 0.7614
             Prevalence : 0.4224
         Detection Rate : 0.3053
   Detection Prevalence : 0.5093
      Balanced Accuracy : 0.6848

       'Positive' Class : 1
```

Model : Logistic Regression

```
> exp(coefficients(Model))
     (Intercept)  survival_1_year1        n_scoreN1        n_scoreNX        tumor_rate          rd_thrpy1
    2.023439e-08      2.736751e+07     3.664889e-01     9.334564e-01      9.950648e-01      6.243071e-01
      m_scoreM1a        m_scoreM1b       m_scoreM1c  gleason_scrLow gleason_scrMedium              U051
    4.338324e-01      3.765116e-01     4.683744e-01     1.823807e+00      1.400215e+00      6.438194e-01
        rad_rem1              S101      brch_thrpy1          stageIIA          stageIIB         stageIII
    6.873509e-01      6.255635e-01     7.372100e-01     1.062790e+00      9.452426e-01      1.107958e+00
         stageIV    Total_Therapy2   Total_Therapy3   Total_Therapy4   Total_Therapy5   Total_Therapy6
    8.471943e-01      8.859146e-01     1.095476e+00     9.590778e-01      1.101605e+00      6.234382e-08
      cry_thrpy1     age_grpGroup2    age_grpGroup5           smoker1             race2             race3
    8.589825e-01      1.981075e+00     1.823207e+00     1.261598e+00      1.314776e+00      1.395156e+00
            race4          psa_rate
    1.281420e+00      1.001107e+00
```

- On average, Patients with low gleason_score are 82% more likely to survive 7 years after diagnosis compared to patients with high gleason_score
- On average, Patients with ‘M1b’ m_score are 63% less likely to survive 7 years after diagnosis compared to patients with ‘M0’ m_score

Results

	Accuracy	Sensitivity	Specificity
Logistic Regression	68%	72%	66%