1 MARCH 2022 UIC IDS 506

PRATHAMESH BAPAT

PROSTATE CANCER SURVIVAL ANALYSIS

Table of Contents

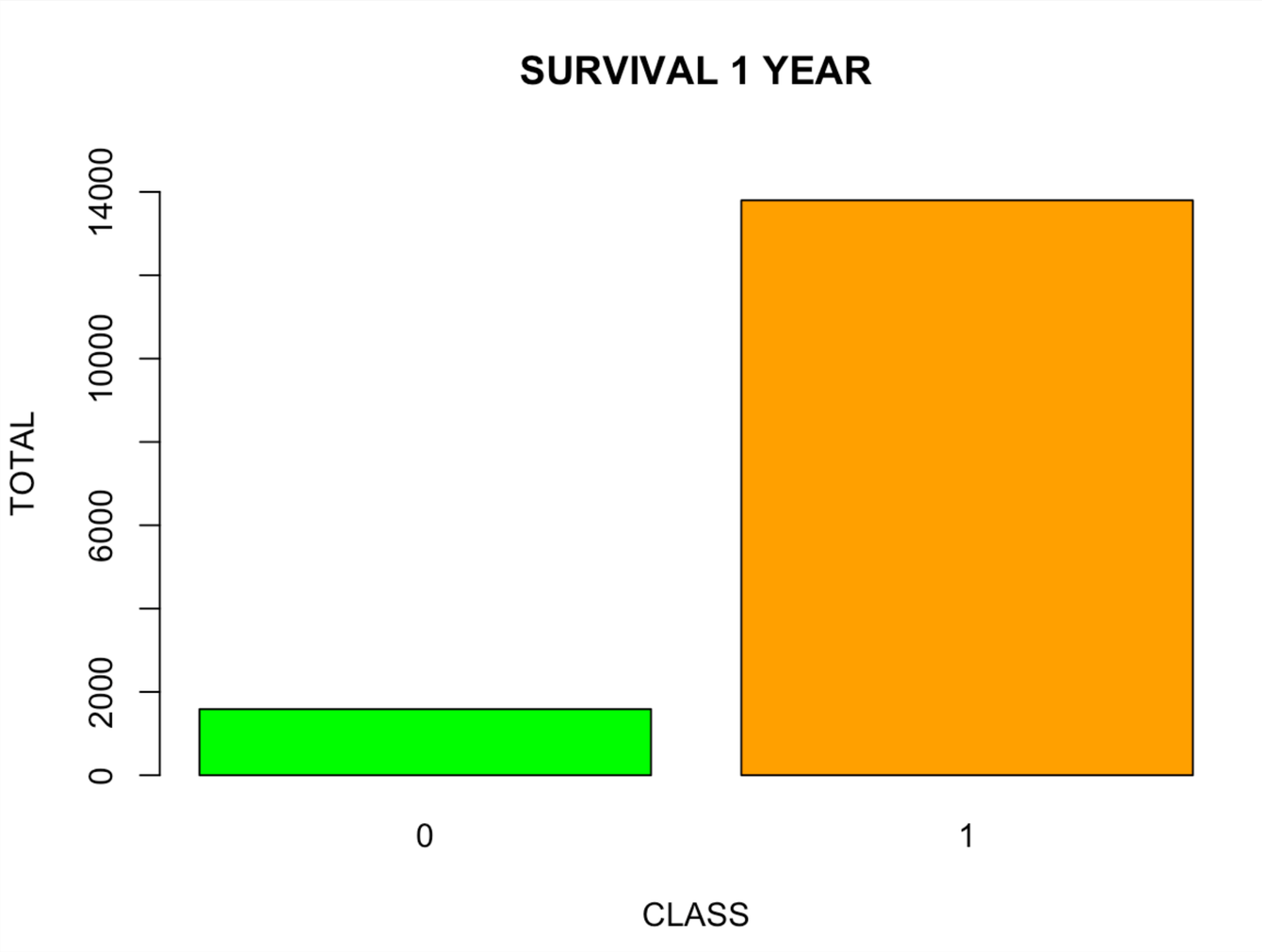
		Page
I	Understanding the Problem	3
II	Grouping features for better analysis	4
III	Feature Selection	6
IV	Logistic Regression	12
V	Conclusion & Step-Wise Regression	14

OBJECTIVE

You are trying to determine the 7-year survival of prostate cancer patients. A patient survived if they are still alive 7 years after diagnosis. This means that a patient is counted as dead whether or not the death was due to their cancer. You have been given details about the patients and their cancers to help you with your prediction.

THE TASK IS TO UNDERSTAND WHICH FEATURES ARE IMPORTANT IN CAUSING CANCER AND ALSO WHICH FEATURES PLAY A VITAL ROLE IN CANCER SURVIVAL. AT LAST THESE FEATURES WILL FIT ON A PREDICTIVE MODEL.

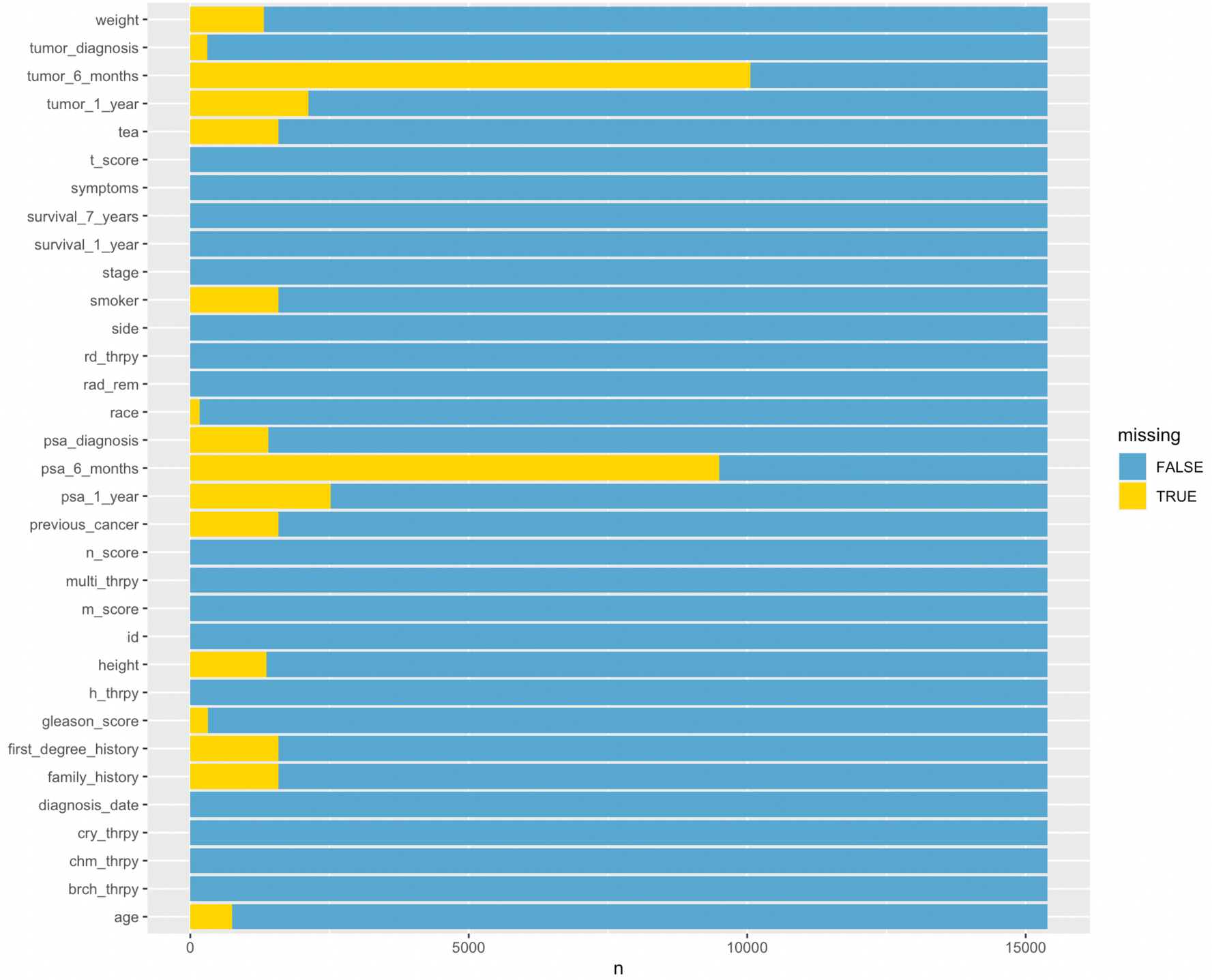
GROUPS
THERAPY
TUMOR RELATED
DEMOGRAPHICS
DIAGNOSIS
SYMPTOMS
OTHER-NON RELATED TO PROSTATE CANCER
NO RELATION - (DROP & FILTER)
SURVIVAL 7 YEARS (TARGET)



SURVIVAL 1 YEAR

Drop the rows with value 0

FINAL ROW SIZE = 13799



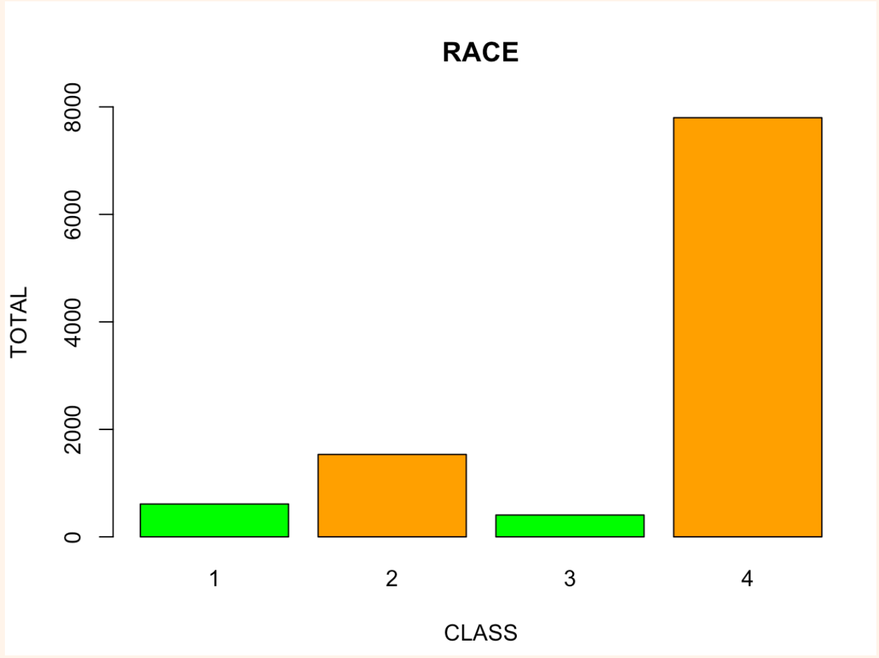
MISSING VALUES AND NULL VALUES

TARGET

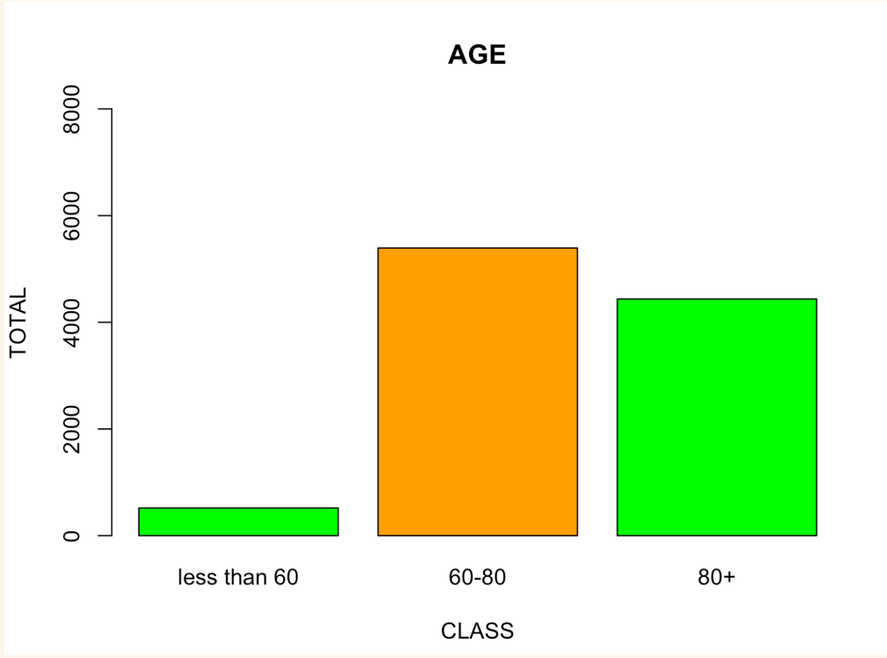
DELETE FEATURES
id
diagnosis_date
tea
previous_cancer
side
survival_1_year
family_history
first_degree_history

Demographics

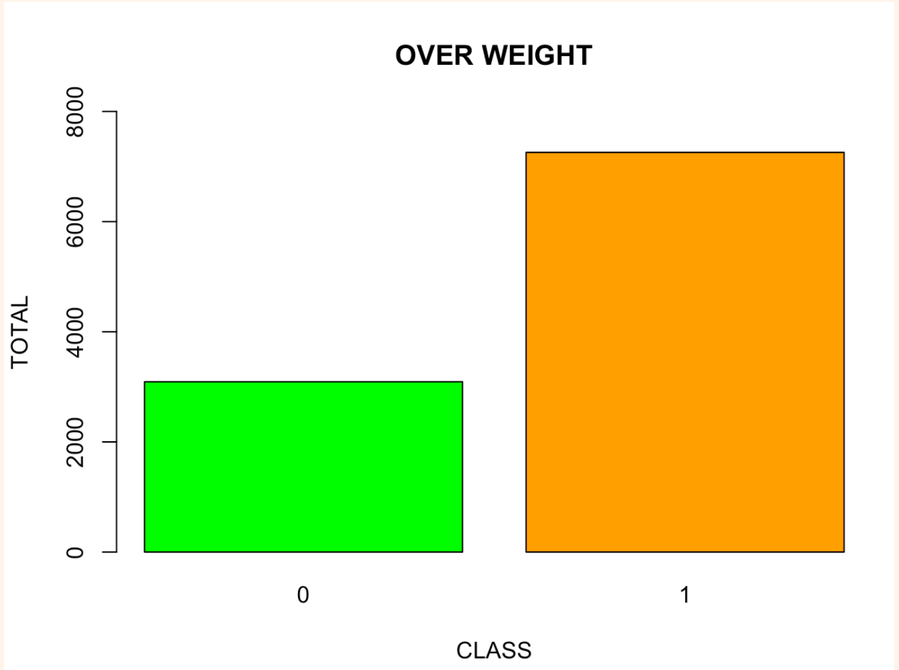
- age
- race
- height
- weight



REPLACE NULLS WITH MODE



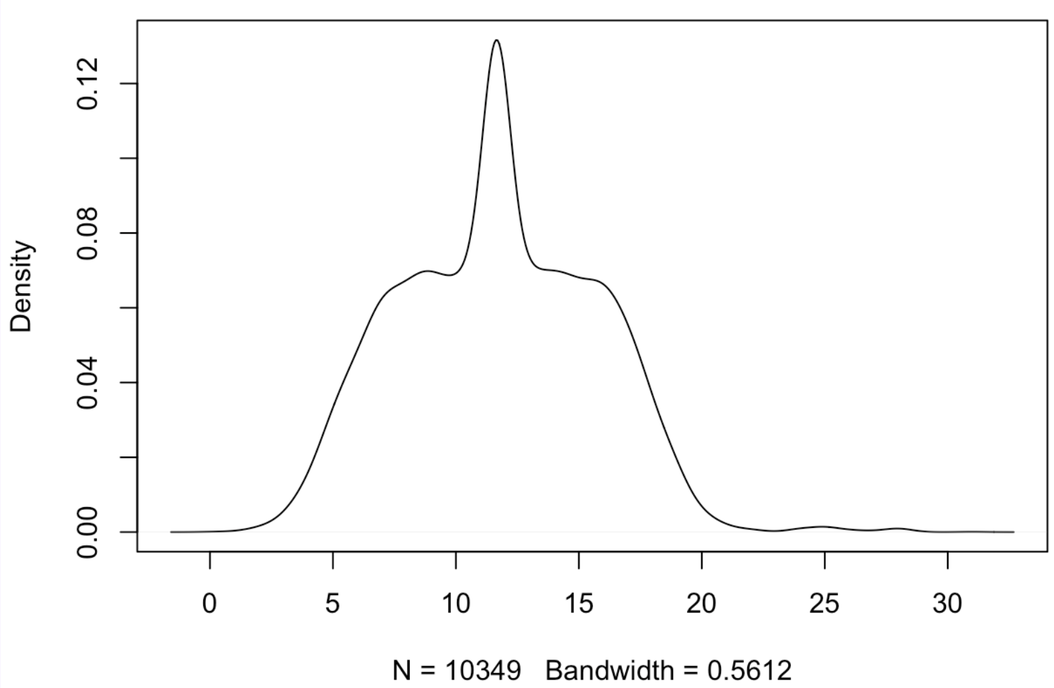
REPLACE NULLS WITH MEAN



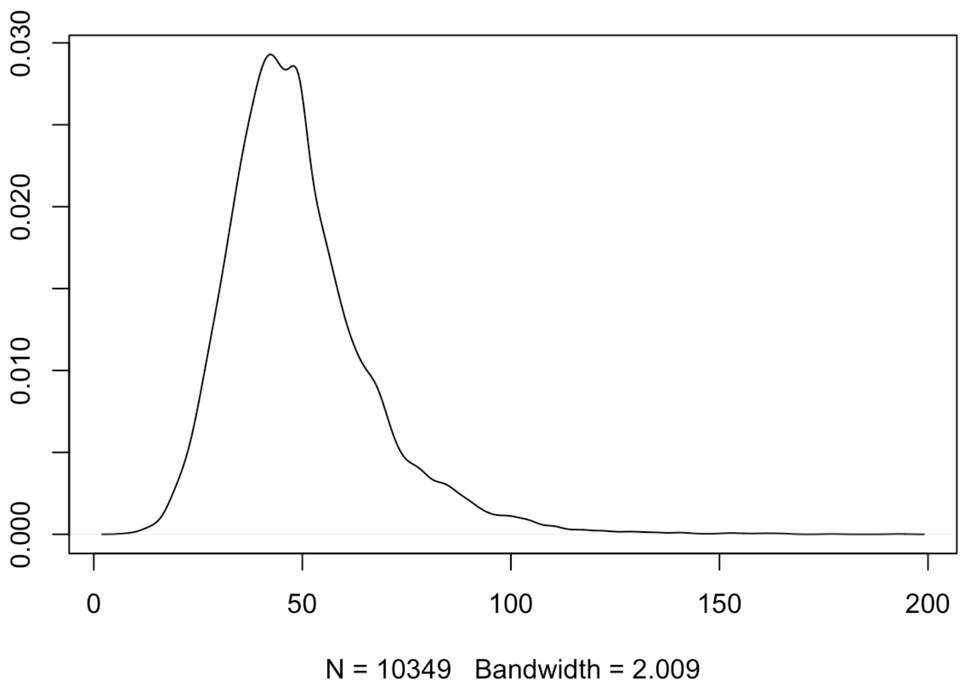
CREATE BMI FROM WEIGHT & HEIGHT

DIVIDE BMI IN TWO PARTS AT THRESHOLD 25

Diagnosis
tumor_diagnosis
tumor_6_months
tumor_1_year
psa_diagnosis
psa_6_months
psa_1_year



psa diagnosis - NORMALIZE



tumor diagnosis - NORMALIZE

```
data:  df$tumor_diagnosis and df$tumor_1_year
S = 8.0436e+10, p-value < 2.2e-16
```

T test - DROP tumor 1 year

```
data:  train$psa_diagnosis and train$psa_1_year
S = 3.1408e+10, p-value < 2.2e-16
```

T test - DROP PSA 1 year

Therapy	WHAT TO DO
multi_thrpy	**Binary - Dependency with other therapies. Drop this**
rd_thrpy	Binary - Factor it
h_thrpy	Binary - No dependency with target. Drop this
chm_thrpy	Binary - Factor it
cry_thrpy	Binary - Factor it
brch_thrpy	Binary - Factor it
rad_rem	Binary - No dependency with target. Drop this

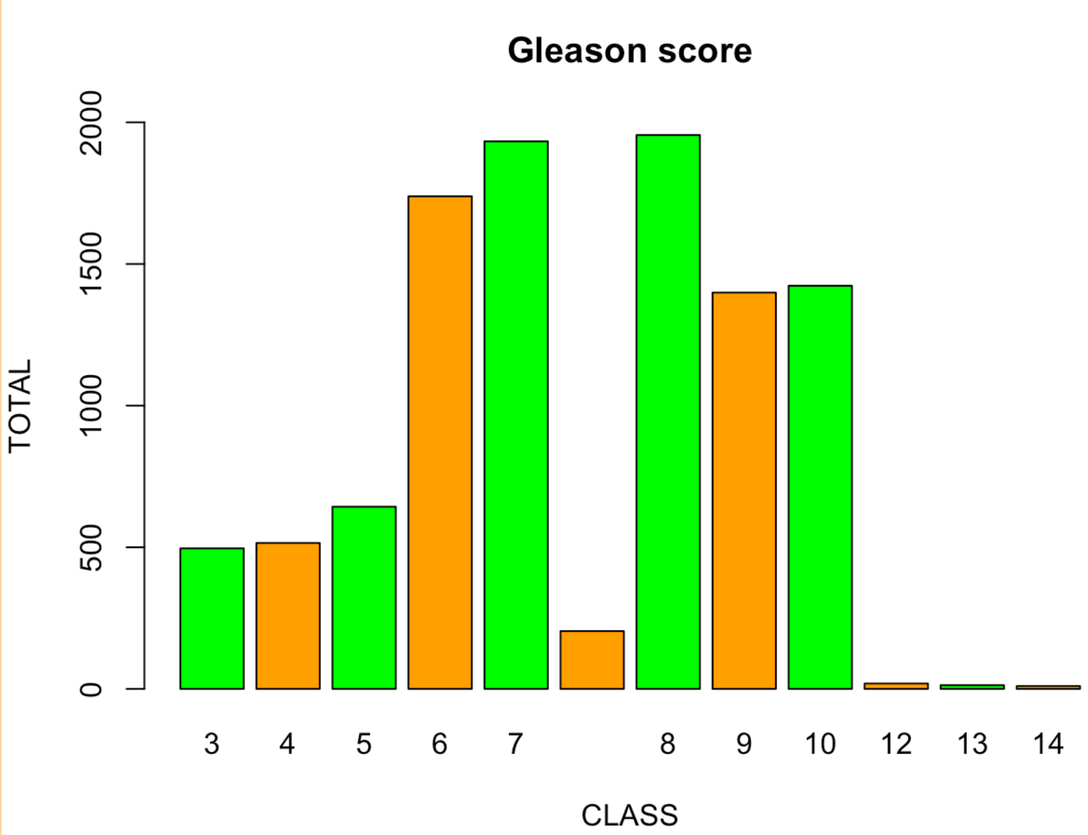
```
data:  tbl4
X-squared = 0.82759, df = 1, p-value = 0.363
```

CHI_SQR TEST - DROP rad_rem

```
data:  tbl4
X-squared = 2.5504, df = 1, p-value = 0.1103
```

CHI_SQR TEST - DROP h_thrpy

TUMOR RELATED	WHAT TO DO
gleason_score	<8 = 0 and >8 =1. Then factor those values (gscat)
t_score	Used to find stage and highly related to stage(chi square test)
n_score	Used to find stage and highly related to stage(chi square test)
m_score	Used to find stage and highly related to stage(chi square test)
stage	Factor the 5 values



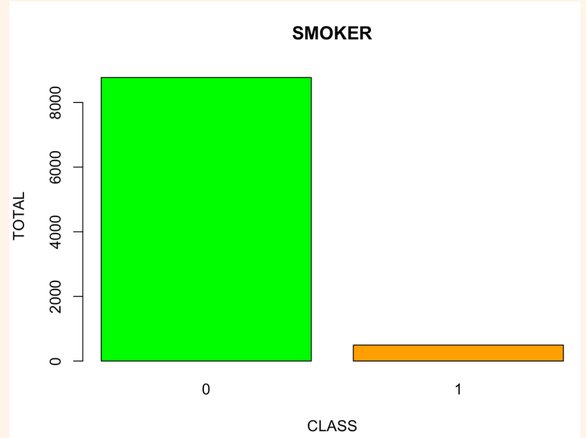
GLEASON SCORE

STAGE
DECIDED TAKING INTO ACCOUNT
t, n and m SCORES

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.03599    0.04212   0.854   0.3929
symptoms_O011 -0.11421    0.17579  -0.650   0.5159
symptoms_O081 -0.48045    0.22242  -2.160   0.0308 *
symptoms_O091 -0.62001    0.27108  -2.287   0.0222 *
symptoms_O101 -0.25772    0.26111  -0.987   0.3236
symptoms_O111  0.03851    0.04822   0.799   0.4245
symptoms_P011 -1.08800    0.13280  -8.193 2.55e-16 ***
symptoms_P021 -1.22092    0.15953  -7.653 1.96e-14 ***
symptoms_P031 -1.84392    0.31438  -5.865 4.49e-09 ***
symptoms_S041  0.03083    0.04001   0.771   0.4410
symptoms_S071  0.06082    0.03533   1.721   0.0852 .
symptoms_S101 -0.58318    0.07854  -7.425 1.13e-13 ***
symptoms_U011  0.03231    0.03543   0.912   0.3618
symptoms_U021 -0.04079    0.03478  -1.173   0.2409
symptoms_U031  0.01430    0.03637   0.393   0.6941
symptoms_U051 -0.44476    0.05853  -7.599 2.98e-14 ***
symptoms_U061  0.02209    0.04303   0.513   0.6076
---
```

LOGISTIC REG WITH JUST THE SYMPTOMS

	1	Multi Label Binarizer – Create seperate features for all 17 symptoms
symptoms	2	**Significant features after LR fit – 7 symptoms.**
	3	**Only two balanced – O11 and U05**

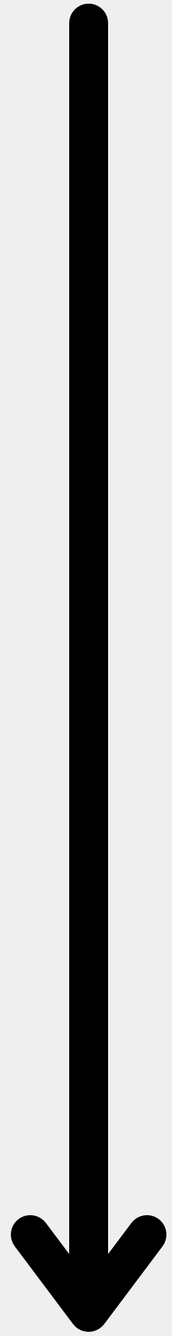


EXTREMELY UNBALANCED - DROP

LOGISTIC REGRESSION

- **SPLIT THE DATA - 75:25**
- **RUN LOG REG with 12 SELECTED FEATURES on TRAIN SET**

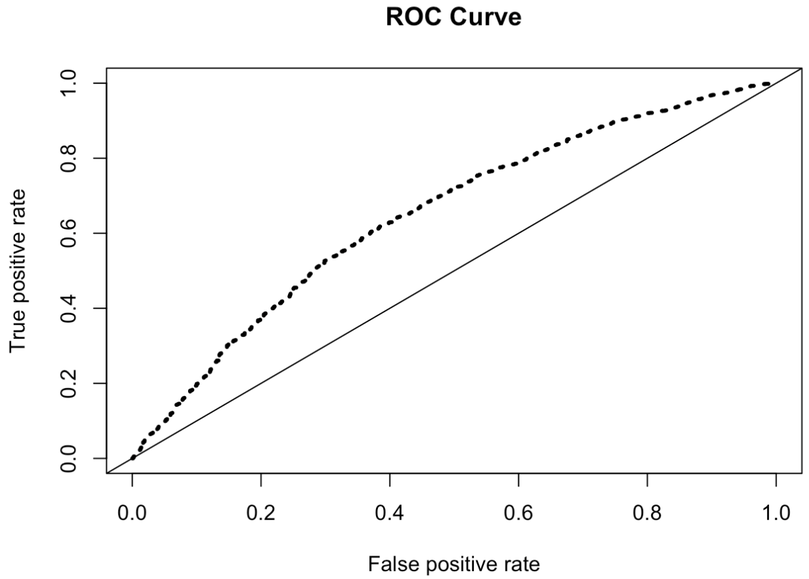
```
model = glm(formula = survival_7_years ~ rd_thrpy+
chm_thrpy+
cry_thrpy+
brch_thrpy+
gscat+
stage+agecat+
race+
tumor_diagnosis+
psa_diagnosis+
symptoms_O11+
symptoms_U05,
data = train, family="binomial")
```



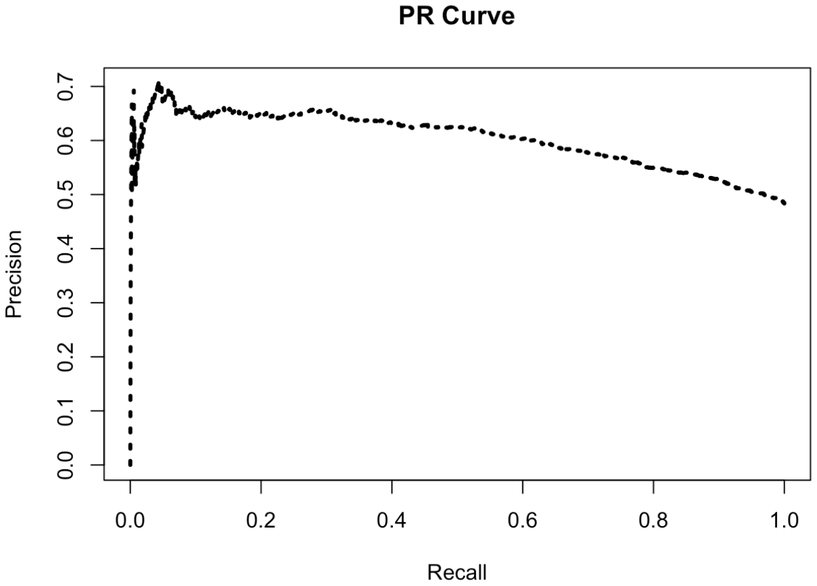
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.153796	0.180679	6.386	1.70e-10	***
rd_thrpy1	-0.321109	0.042583	-7.541	4.67e-14	***
chm_thrpy1	-0.001764	0.051413	-0.034	0.972625	
cry_thrpy1	-0.050575	0.052025	-0.972	0.330995	
brch_thrpy1	-0.153452	0.052510	-2.922	0.003474	**
gscat8+	-0.479827	0.048552	-9.883	< 2e-16	***
stageIIA	-0.305608	0.128236	-2.383	0.017164	*
stageIIB	-0.445233	0.128043	-3.477	0.000507	***
stageIII	-0.249067	0.129465	-1.924	0.054378	.
stageIV	-1.102263	0.128054	-8.608	< 2e-16	***
agecat60-80	0.176226	0.096549	1.825	0.067965	.
agecat80+	-0.001900	0.097359	-0.020	0.984428	
race2	0.307054	0.100094	3.068	0.002157	**
race3	0.304304	0.133671	2.277	0.022815	*
race4	0.229773	0.088023	2.610	0.009045	**
tumor_diagnosis	-0.006080	0.001210	-5.027	4.99e-07	***
psa_diagnosis	-0.010333	0.005433	-1.902	0.057176	.
symptoms_0111	0.038566	0.056402	0.684	0.494128	
symptoms_U051	-0.460702	0.069614	-6.618	3.64e-11	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



ACCURACY = 62%



```r
set.seed(123)
# Set up repeated k-fold cross-validation
library(caret)
train.control <- trainControl(method = "cv", number = 10)
# Train the model
step.model <- train(survival_7_years ~., data = train_1,
                    method = "leapBackward",
                    tuneGrid = data.frame(nvmax = 1:12),
                    trControl = train.control
)
step.model$results
```

BACKWARD STEPWISE REGRESSION

GET THE BEST MODEL(BEST R-SQAURED)

WHAT ELSE CAN WE DO?

```
> model$results
   nvmax      RMSE   Rsquared       MAE      RMSESD  RsquaredSD       MAESD
1      1 0.4910345 0.03480620 0.4821589 0.001614610 0.007379619 0.001446058
2      2 0.4858125 0.05537317 0.4719209 0.002464711 0.009638606 0.002172849
3      3 0.4844959 0.06051746 0.4692989 0.002548783 0.010147151 0.002128610
4      4 0.4834029 0.06471289 0.4671366 0.002568669 0.010013533 0.002050140
5      5 0.4828750 0.06673317 0.4660410 0.002596250 0.010499414 0.002070519
6      6 0.4824666 0.06837736 0.4652216 0.002774394 0.011227919 0.002200821
7      7 0.4824707 0.06836629 0.4649738 0.002729411 0.011245807 0.002173278
8      8 0.4822923 0.06913553 0.4645905 0.002815916 0.011311598 0.002150740
9      9 0.4823713 0.06882530 0.4645805 0.002827891 0.011375317 0.002201031
10    10 0.4822855 0.06911997 0.4644471 0.002751759 0.011122397 0.002114233
11    11 0.4822831 0.06913758 0.4643402 0.002684042 0.010784956 0.002021506
12    12 0.4821124 0.06978440 0.4640966 0.002756647 0.011092277 0.002119270
```

Thank you!