SSLC Data Analysis

Group 27

MT2014081 Nupur Garg
MT2014085 Setu Patani
MT2014089 Praveen Baby

Outline

- Data Preparation
- Exploratory Analysis
- Classification
- Regression
- Association Rule Mining
- Clustering

Data Preparation

- Removed NA values
- Replaced 888 with 0
- Removed * from Values
- Scaled L1 marks and Made respective changes to Total Marks and NRC Class

Karnataka

	Count
District	34
Taluq	204
School	9800
Center	3001

Average No of Students/School

Average No of Students/School

District Name	Student/School
KODAGU	2.52
UTTARA KANNADA	2.77
GADAG	2.72
KOPPAL	2.83
UDUPI	2.85

TOP 5

Average : 2.96

District Name	Student/School
HASSAN	3.43
DAKSHIN KANNADA	3.34
CHAMARAJANAGAR	3.23
BANGALORE URBAN	3.13
GULBARGA	3.08

Bottom 5

Inference : Districts in Bottom 5 Table needs to open more Schools

Average No of Students/Center

District Name	Student/Center
DAKSHIN KANNADA	12.67
HASSAN	12.01
GULBARGA	11.88
CHAMARAJANAGAR	11.62
BIDAR	10.9

TOP 5

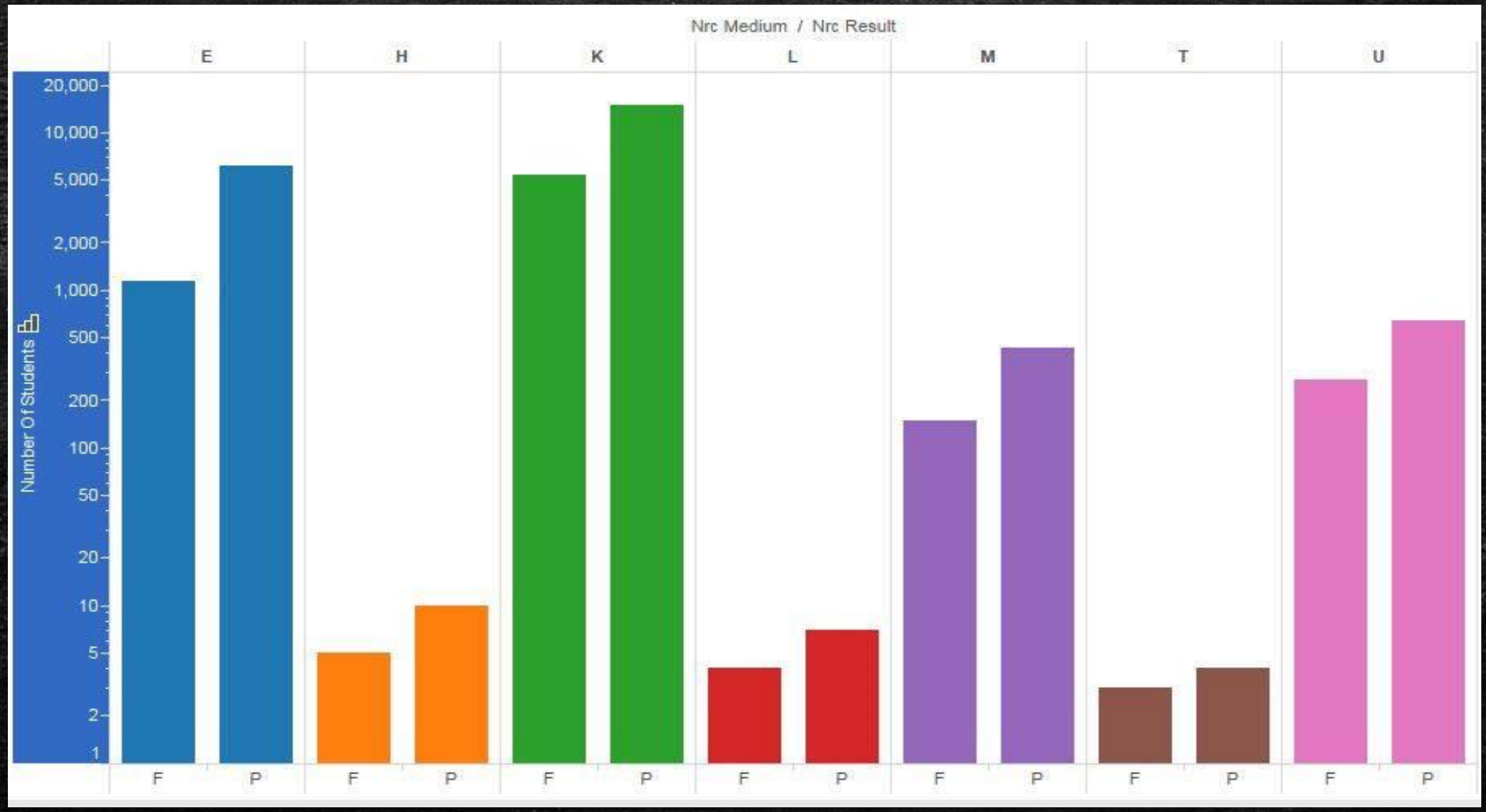
Average : 9.80

District Name	Student/Center
CHITRADURGA	8.9
BANGALORE URBAN	8.99
HAVERI	9.00
MYSORE	9.10
BELLARY	9.13

Bottom 5

Inference : Districts in Top 5 Table can accommodate more Students

School Medium Analysis



School Medium Analysis

Medium	No of Students	No of Schools	Students/Schools
Kannad	20558	7250	2.84
English	7323	2772	2.64
Hindi	15	7	2.14
Marathi	582	225	2.59
Tamil	7	5	1.4
Telugu	11	11	1
Urdu	920	333	2.76

Inference : In SSLC, Kannad Medium has majority of Students while Tamil has the lowest.

District Wise Performance

District Wise Performance

District Name	Pass Percentage
UTTAR KANNADA	84.3
UDUPI	83.38
MANDYA	82.80
BELGAUM	82.73
DAKSHIN KANNADA	82.14

TOP 5

Average : 76.08 %

District Name	Pass Percentage
BIDAR	51.19846596
GULBARGA	64.98
BIJAPUR	69.29057337
KOLAR	71.32
BANGALORE URBAN	71.41

Bottom 5

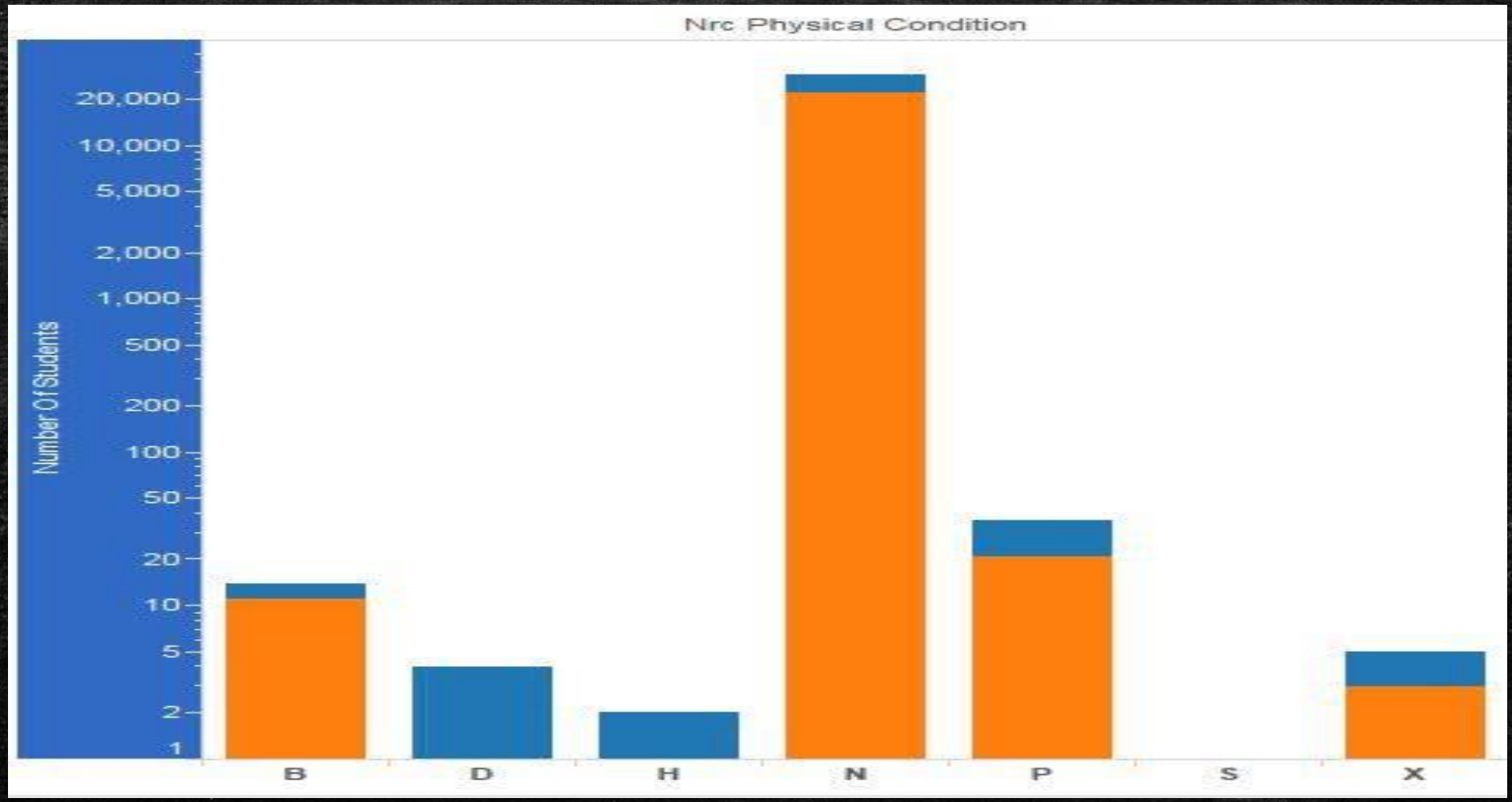
Inference : Districts mentioned in Bottom 5 table needs lot of improvement

School Wise Performance

Percentage Range	Count
100	5827
100-80	448
80-60	1154
60-40	1012
40-1	430
0	929

Inference : Schools With Below 40% Result needs improvement

Physical Condition - Performance



Top Schools For Physically Handicap

School Name
VINAYAKA JUNIOR COLLEGE HAGALAVADI, GUBBI TALUK TUMKUR DISTRICT
THE NEW CAMBRIDGE HIGH SCHOOL VIJAYANAGAR, BANGALORE
GOVERNMENT JR. COLLEGE AMRUTHUR, KUNIGAL TALUK TUMKUR DISTRICT
GOVERNMENT HIGH SCHOOL BANDERWAD, AFZALAPUR TALUK GULBARGA DISTRICT
GOVT EMPRESS GIRLS COMP P.U. COLLEGE TUMKUR TUMKUR DISTRICT

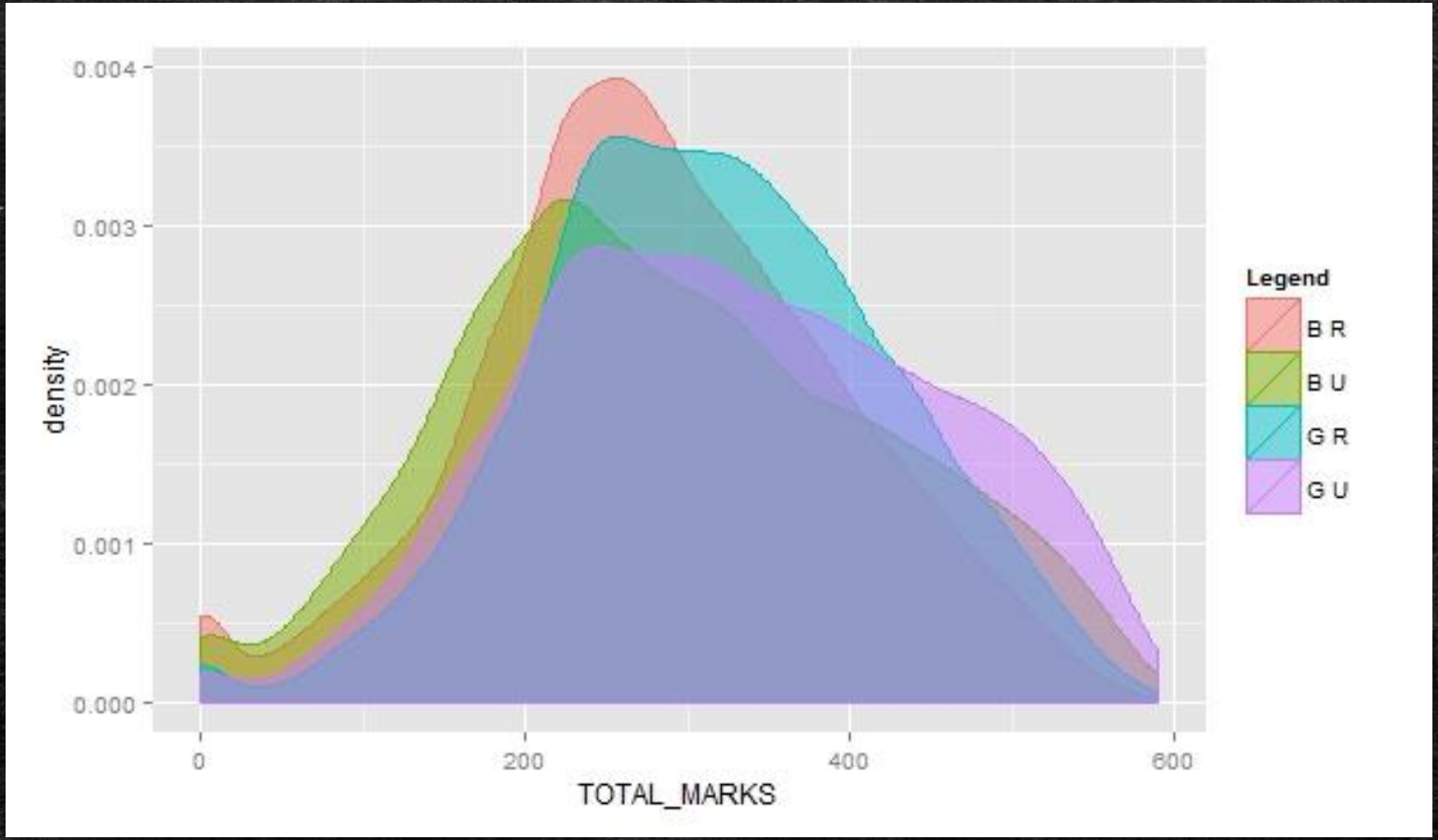
Inference : Top performing School for Physically Handicap Candidate.

Top Schools For Blind Candidate

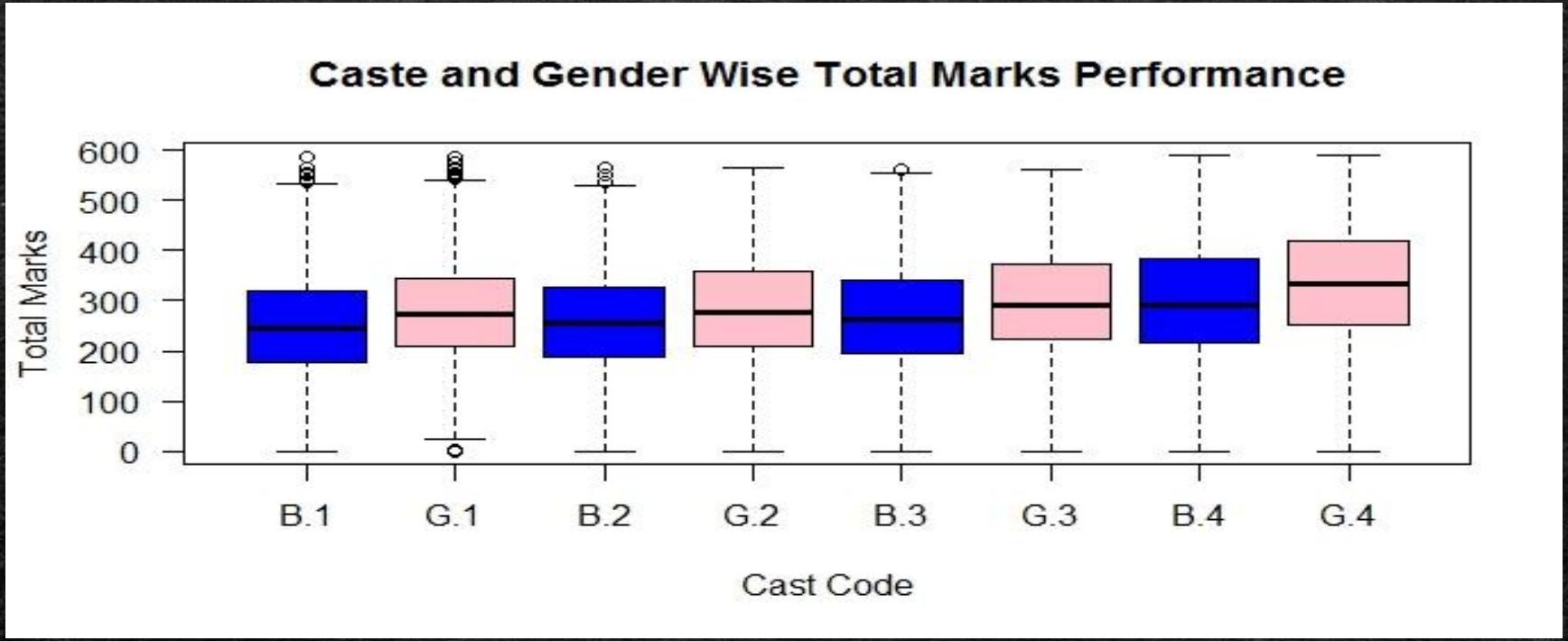
School Name
MAHESHWARI SCHOOL FOR BLIND NEHRUNAGAR BELGAUM DISTRICT
SRI SIDDARUDA SWAMI BLIND CHILDRENS RES HIGH SCHOOL A. V. PUJAR BULDING SIDDARUDA MATADA MAIN ROAD ,OLD HUBLI DHARWAR HUBLI URBAN TQ , DHARWAR DIST.
GOVT.HIGH SCHOOL NAYANADU BANTWAL TALUK D K DISTRICT
J S B FREE RESIDENTIAL SCHOOL FOR BLIND ARCHAKARAHALLI, RAMANAGARAM TALUK RAMANAGARA DISTRICT
GOVERNMENT HIGH SCHOOL TURAKARASHIGIHALLI BAILAHONGAL TQ BELGAUM DISTRICT

Inference : Top performing School for Blind Candidate.

Gender – Urban/Rural Analysis

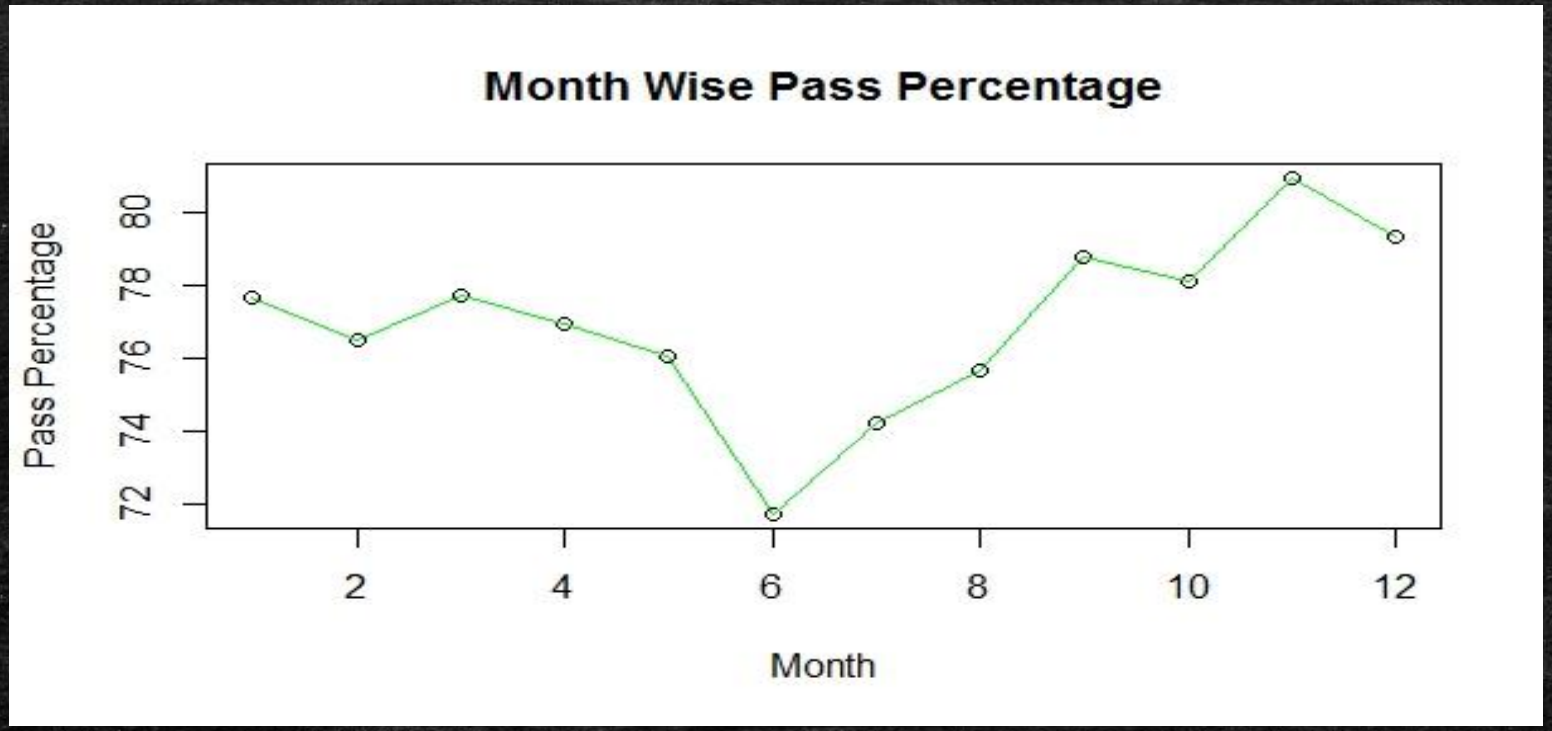


Caste – Gender Performance Analysis



Inference : Girl's performance is better than Boy's in every caste.
Caste 4 Performance is best.

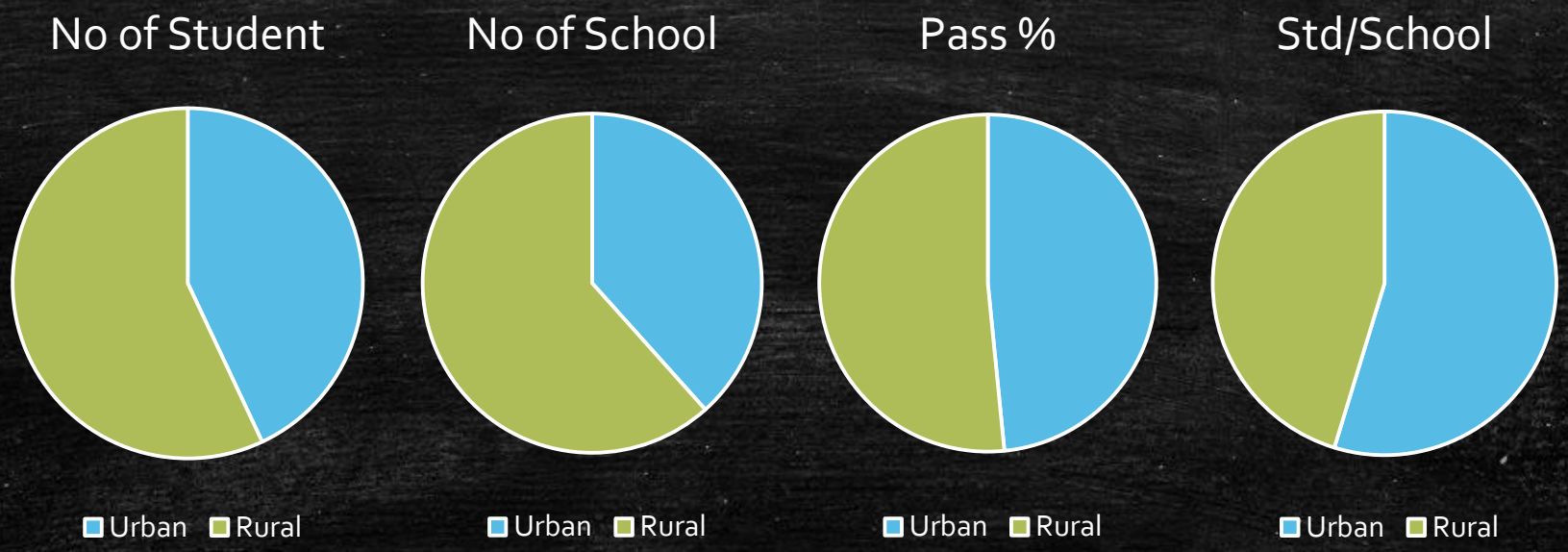
Result based on Months



Inference : Students born in November performs well.

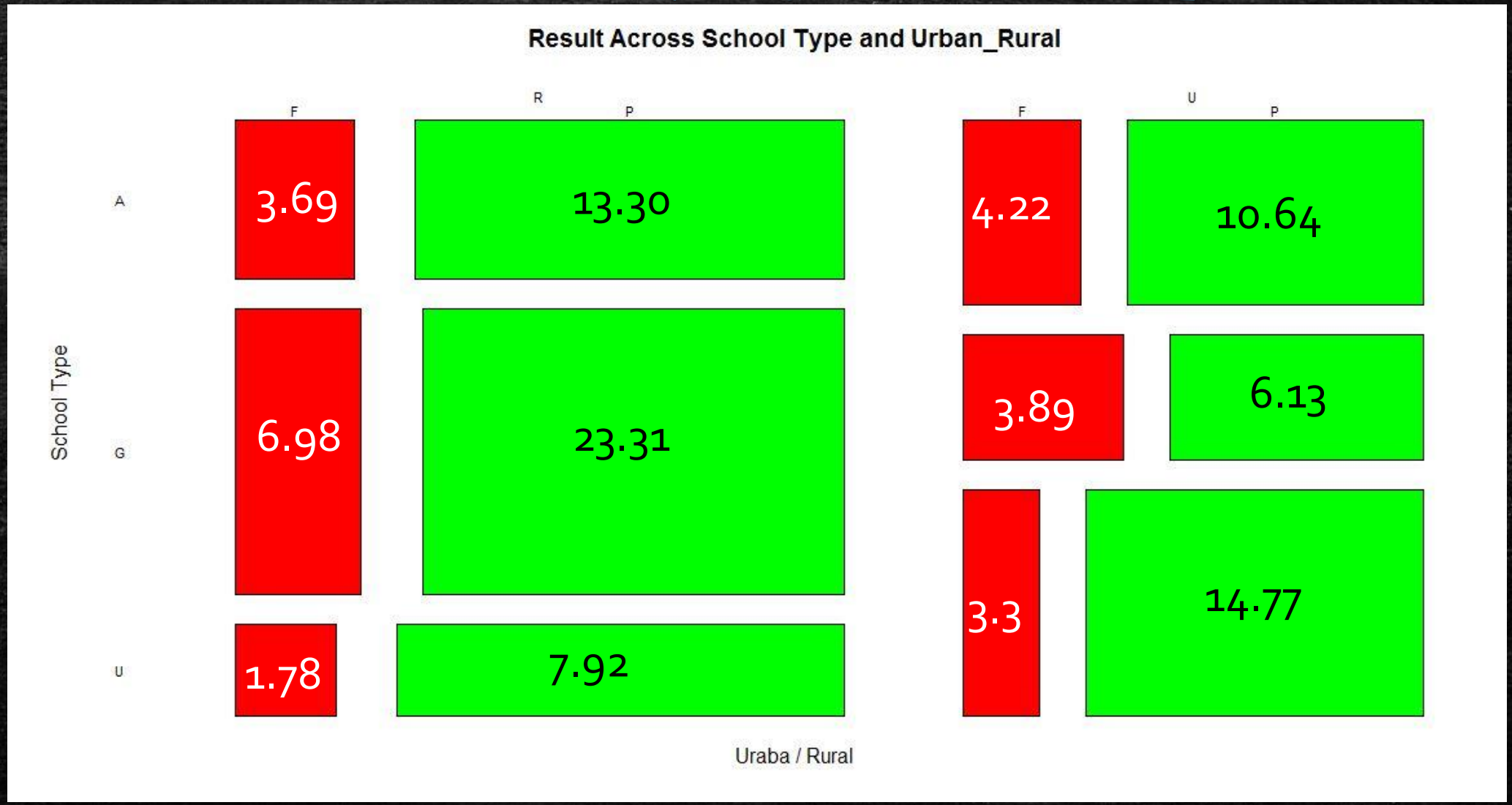
URBAN – RURAL Analysis

	No of Student	No of Schools	Pass %	Students/School
Urban	12651 (43%)	3761 (38.77%)	73.36%	3.36
Rural	16765 (57%)	6039 (61.62%)	78.12 %	2.78



Inference : Relative Area of All Pie chats should be similar

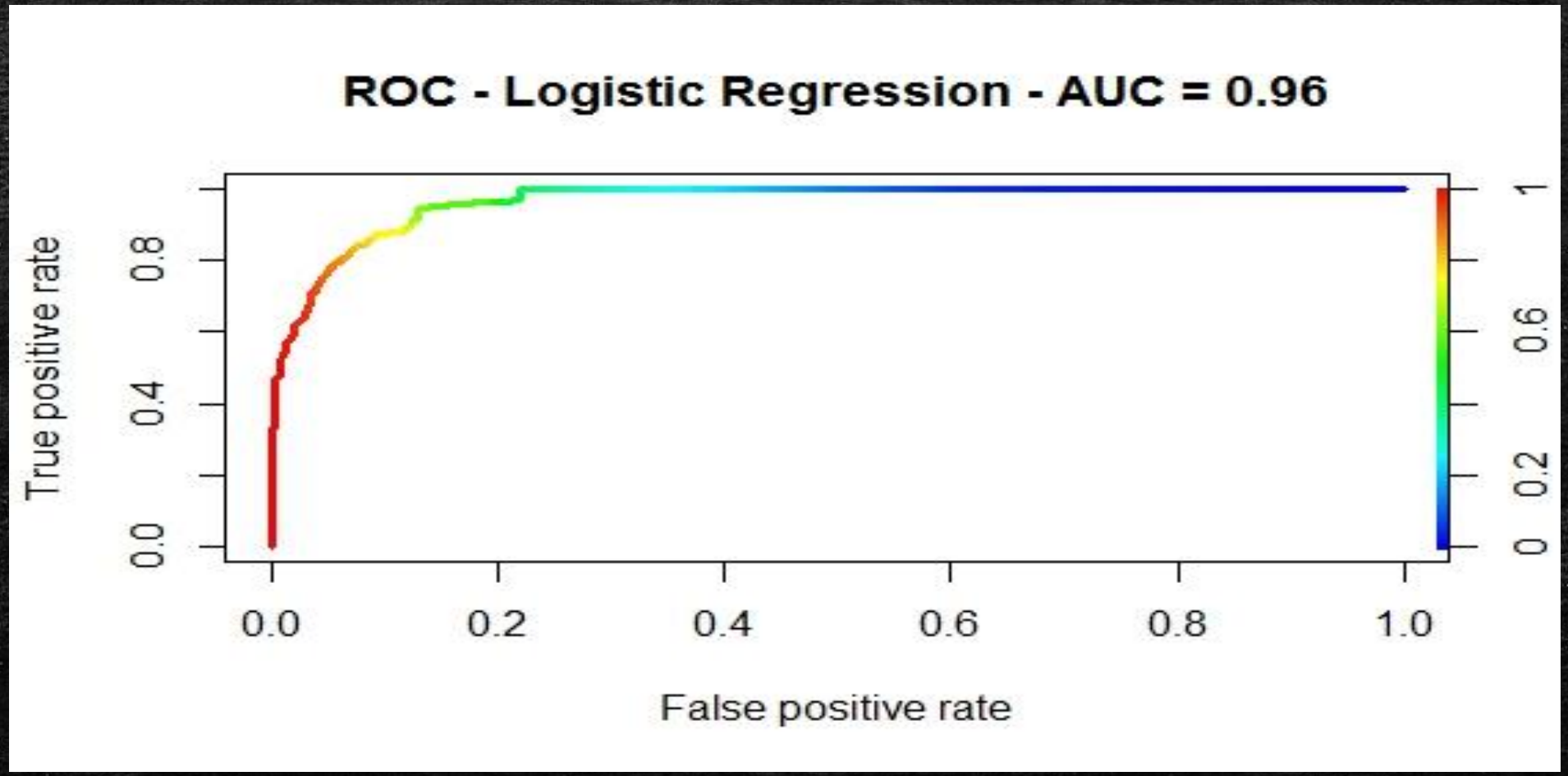
URBAN – RURAL Analysis



Classification

- NRC_RESULT based on S2 and L2 Marks
- ROC curve for Logistic Regression
- Applied Various Techniques like,
  - SVM
  - Random Forest (Cross Validation)
  - Decision Tree
  - LDA
  - Naive Bayes
- Comparison of Accuracy of above techniques.
- NRC_MEDIUM base on L1-L2-L3 Codes

ROC Curve – Logistic Regression



SVM vs RF(CV) vs DT vs LDA vs NB

SVM	F	P
F	802	168
P	0	5446

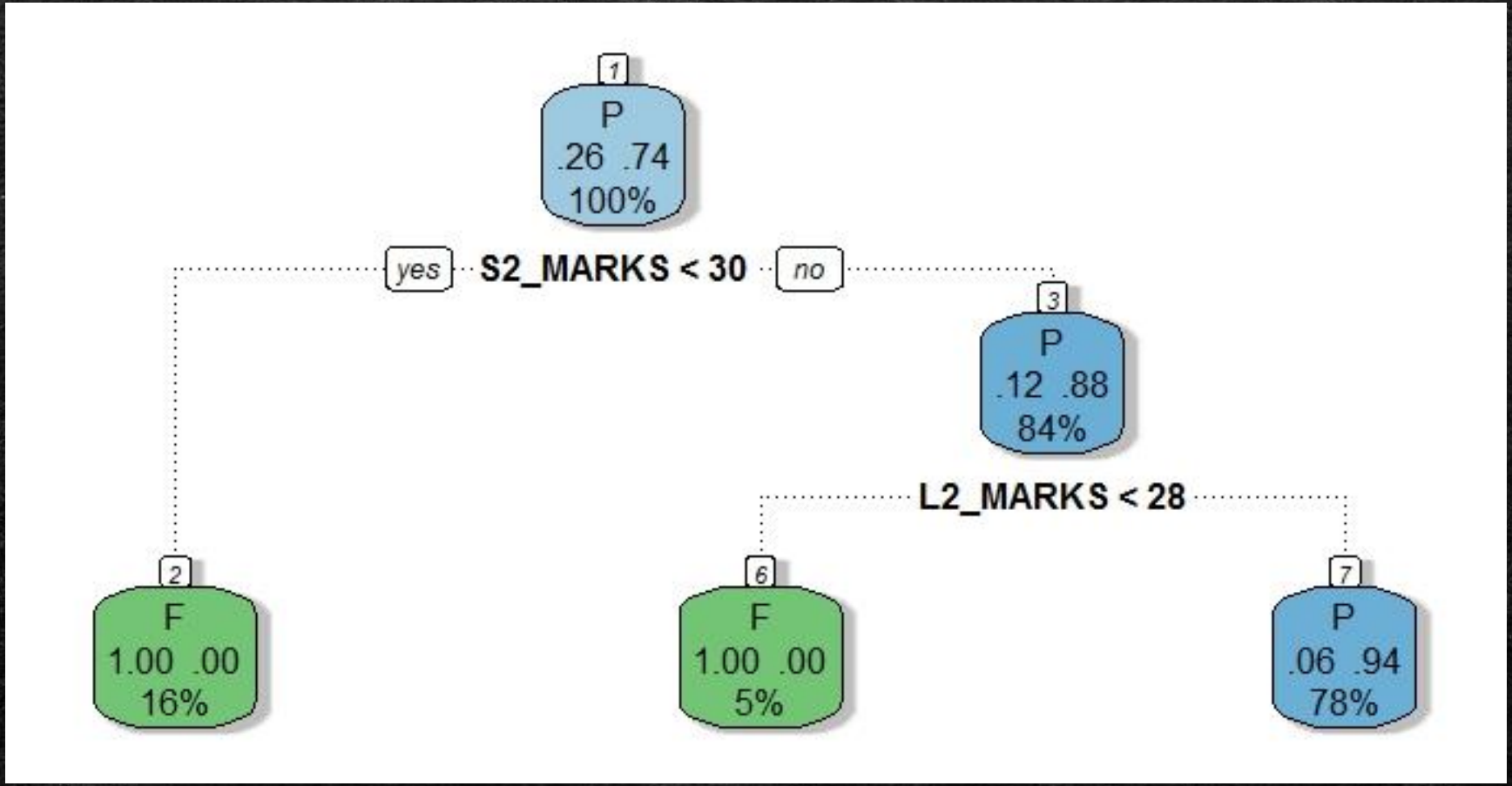
RF	F	P
F	810	160
P	7	5439

DT	F	P
F	810	160
P	0	5446

LDA	F	P
F	725	245
P	0	5446

NB	F	P
F	804	166
P	0	5446

Decision Tree

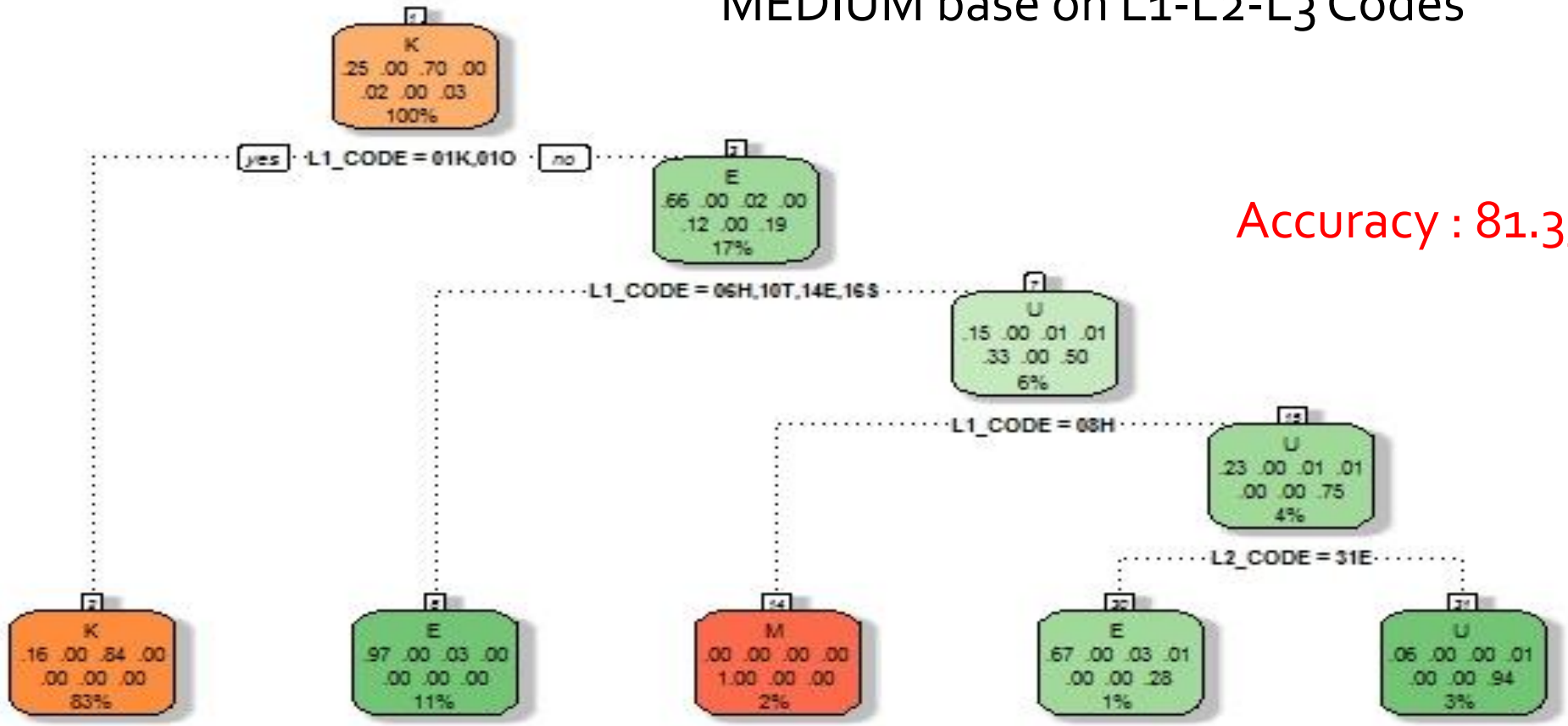


Classification Accuracy

Technique	Accuracy
SVM	97.4 %
Random Forest	97.4 %
Decision Tree	**97.51 %**
LDA	96.18 %
Naive Bayes	97.41 %

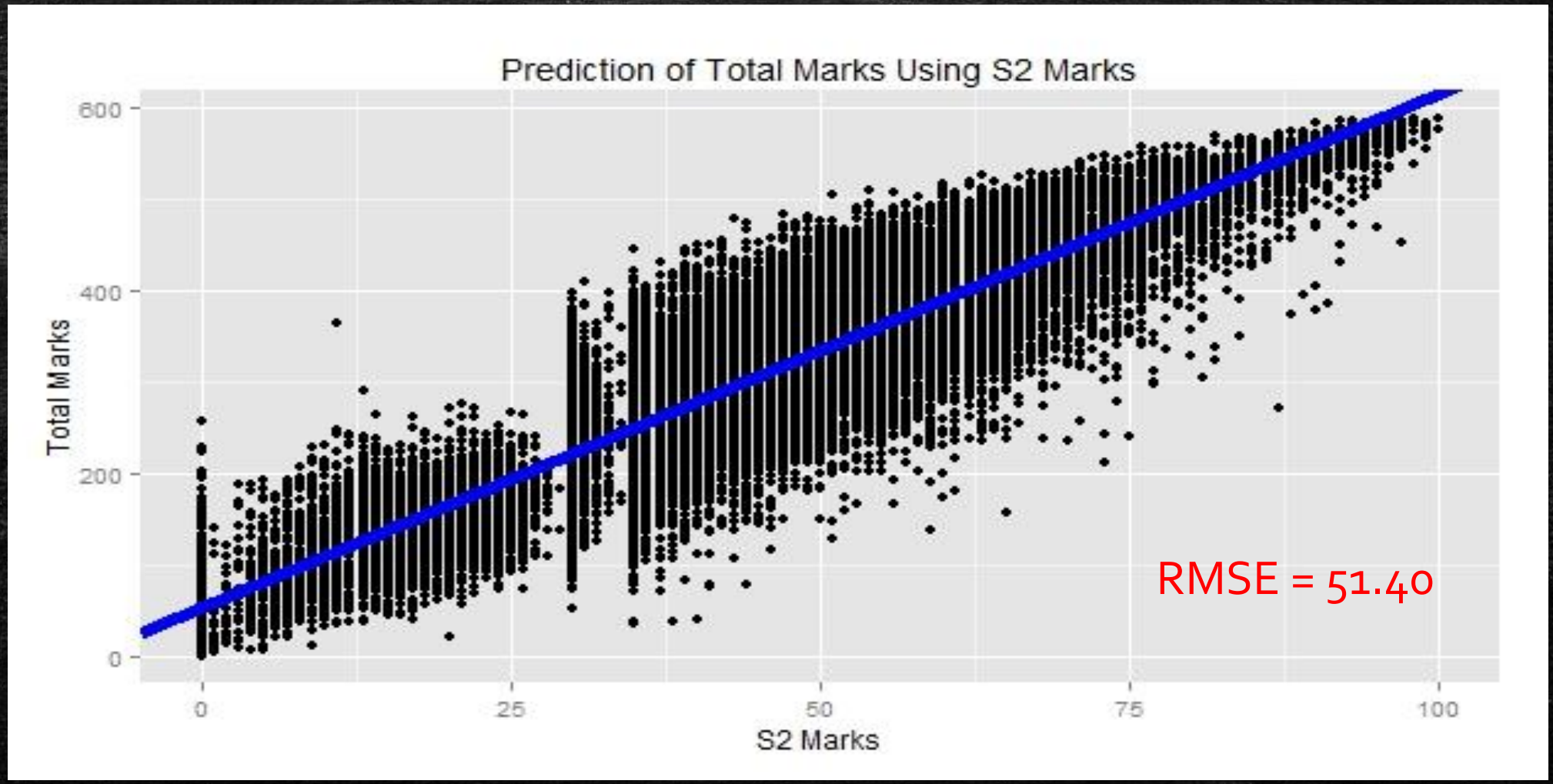
Classification

MEDIUM base on L1-L2-L3 Codes

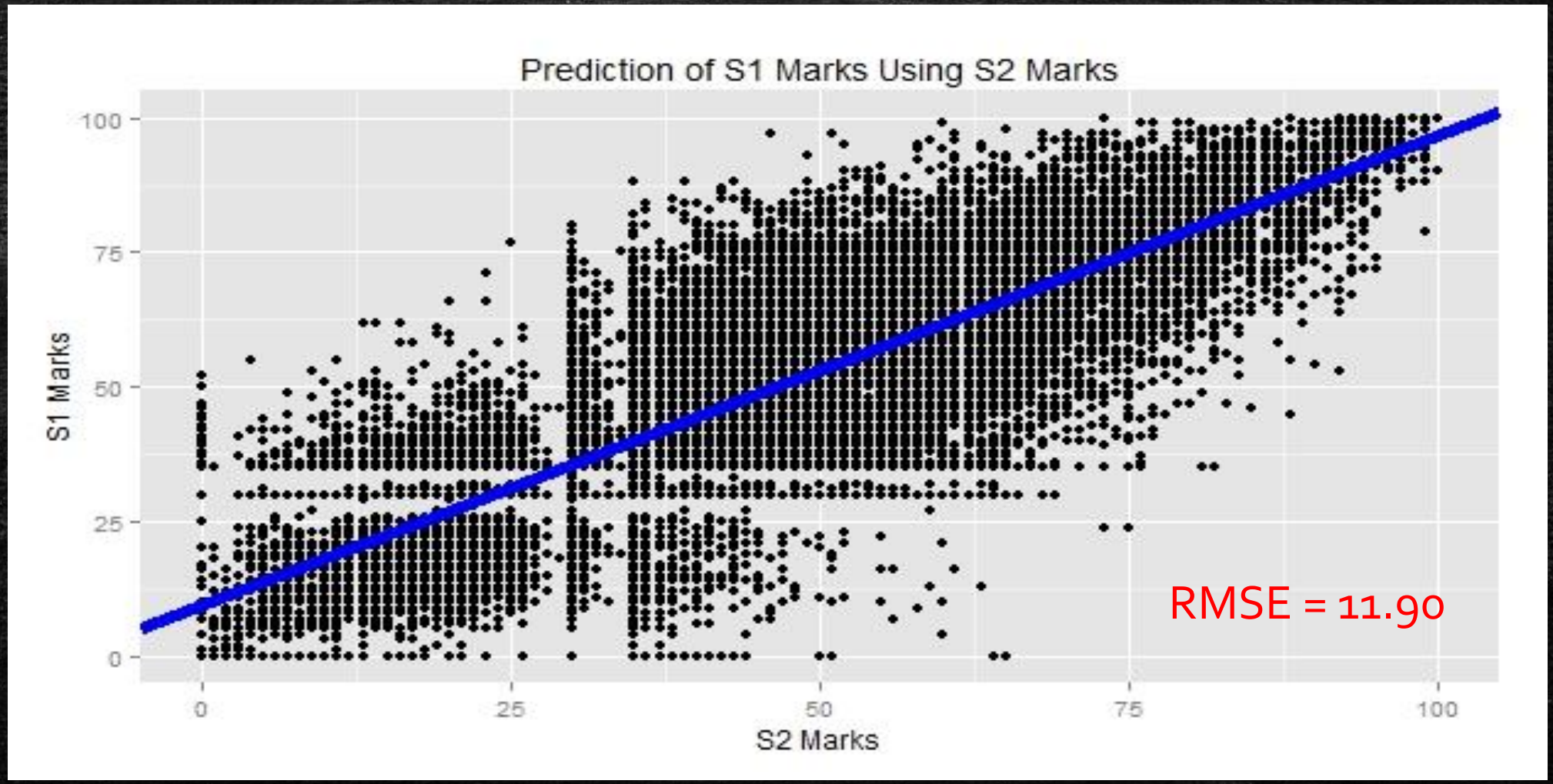


Accuracy : 81.33%

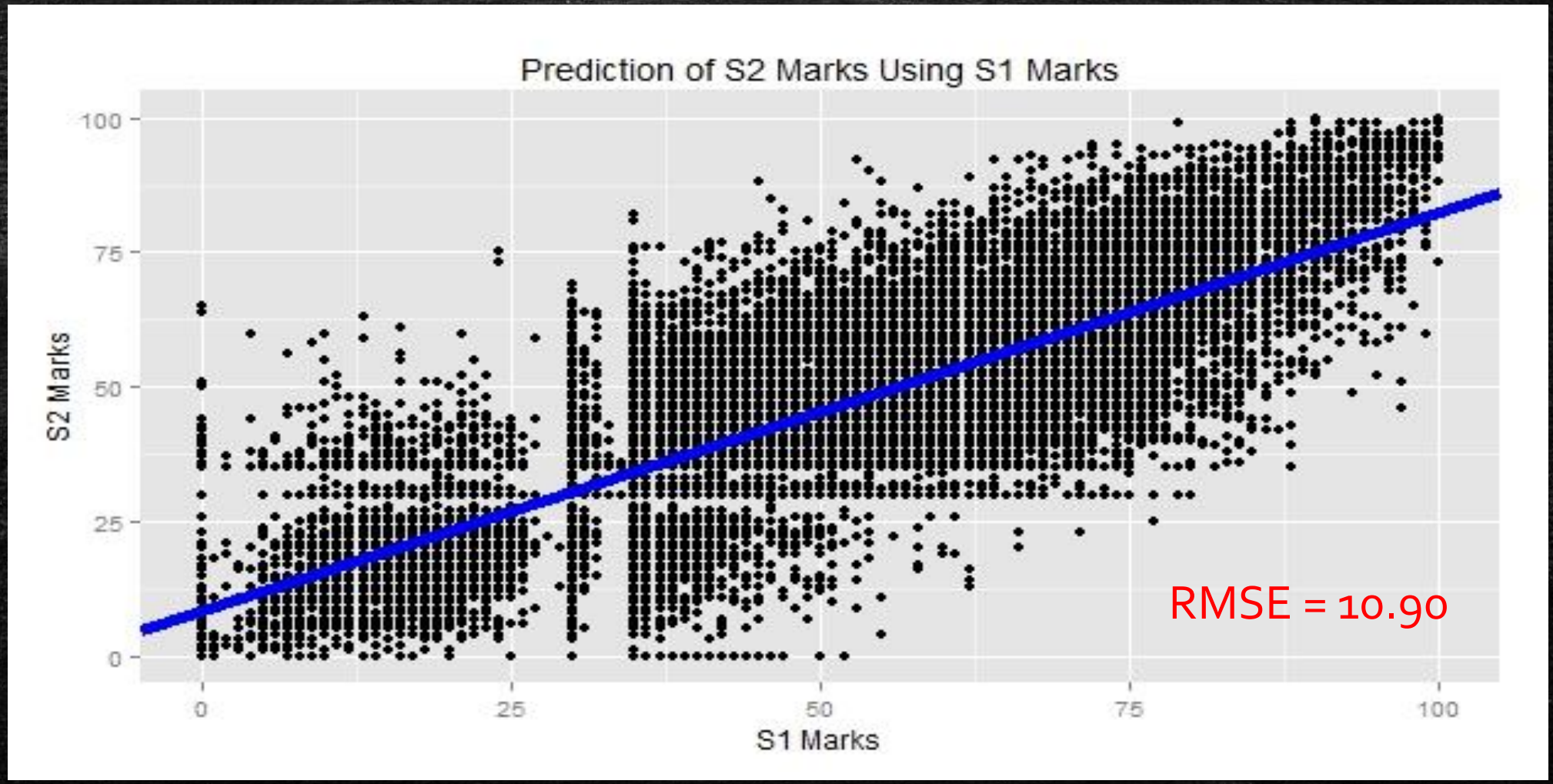
Regression


Prediction of Total Marks Using S2 Marks
Total Marks
600
400
200
0
0
25
50
75
100
S2 Marks
RMSE = 51.40

Regression


Prediction of S1 Marks Using S2 Marks
S1 Marks
100
75
50
25
0
0
25
50
75
100
S2 Marks
RMSE = 11.90

Regression


Prediction of S2 Marks Using S1 Marks
S2 Marks
100
75
50
25
0
0
25
50
75
100
S1 Marks
RMSE = 10.90

Association Rule Mining

	lhs		rhs	support	confidence	lift
2	{NRC_CLASS=PASS,NRC_GENDER_CODE=B}	=>	{NRC_MEDIUM=K}	0.1252040	0.8074984	1.155432
1	{NRC_CLASS=PASS}	=>	{NRC_MEDIUM=K}	0.2237558	0.8072112	1.155021

	lhs		rhs	support	confidence	lift
2	{NRC_MEDIUM=E}	=>	{NRC_CASTE_CODE=4}	0.2111776	0.8482862	1.199096
1	{NRC_CASTE_CODE=1}	=>	{NRC_MEDIUM=K}	0.1515842	0.8339256	1.193246

Association Rule Mining

	lhs		rhs	support	confidence	lift
30	{L2_CODE=31E,L3_CODE=61H}	=>	{L1_CODE=01K}	0.8295145	1.0000000	1.1973786
33	{L2_CODE=31E,L3_CODE=61H,NRC_MEDIUM=E}	=>	{L1_CODE=01K}	0.1385301	1.0000000	1.1973786
36	{L2_CODE=31E,L3_CODE=61H,NRC_MEDIUM=K}	=>	{L1_CODE=01K}	0.6909845	1.0000000	1.1973786
23	{L3_CODE=61H,NRC_MEDIUM=K}	=>	{L1_CODE=01K}	0.6909845	0.9999016	1.1972608
24	{L1_CODE=01K,L3_CODE=61H}	=>	{NRC_MEDIUM=K}	0.6909845	0.8329986	1.1919198
27	{L2_CODE=31E,L3_CODE=61H}	=>	{NRC_MEDIUM=K}	0.6909845	0.8329986	1.1919198
37	{L1_CODE=01K,L2_CODE=31E,L3_CODE=61H}	=>	{NRC_MEDIUM=K}	0.6909845	0.8329986	1.1919198
20	{L2_CODE=31E,NRC_MEDIUM=K}	=>	{L1_CODE=01K}	0.6950299	0.9949389	1.1913186

Association Rule Mining

	lhs		rhs	support	confidence	lift
11	{NRC_MEDIUM=K,L2_CODE=31E}	=>	{L1_CODE=01K}	0.6950299	0.9949389	1.1913186
4	{L1_CODE=01K}	=>	{NRC_MEDIUM=K}	0.6950299	0.8322139	1.1907970
12	{L1_CODE=01K,L2_CODE=31E}	=>	{NRC_MEDIUM=K}	0.6950299	0.8322139	1.1907970
3	{NRC_MEDIUM=K}	=>	{L1_CODE=01K}	0.6950299	0.9945034	1.1907970
6	{L1_CODE=01K}	=>	{L2_CODE=31E}	0.8351577	1.0000000	1.1188620

Association Rule Mining

	lhs		rhs	support	confidence	lift
6	{SCHOOL_TYPE=G,URBAN_RURAL=R}	=>	{L1_CODE=01K}	0.2934457	0.9686904	1.159889
3	{SCHOOL_TYPE=G}	=>	{L1_CODE=01K}	0.3787055	0.9393709	1.124783
4	{URBAN_RURAL=R}	=>	{L1_CODE=01K}	0.5334512	0.9359976	1.120744
5	{SCHOOL_TYPE=A,URBAN_RURAL=R}	=>	{L1_CODE=01K}	0.1558336	0.9166167	1.097537
2	{SCHOOL_TYPE=A}	=>	{L1_CODE=01K}	0.2686973	0.8432732	1.009717
1	{}	=>	{L1_CODE=01K}	0.8351577	0.8351577	1.000000

	lhs		rhs	support	confidence	lift
3	{NRC_MEDIUM=E,NRC_GENDER_CODE=G}	=>	{NRC_RESULT=P}	0.1021553	0.8890533	1.168561
1	{NRC_MEDIUM=E}	=>	{NRC_RESULT=P}	0.2098858	0.8430971	1.108157
2	{NRC_GENDER_CODE=G}	=>	{NRC_RESULT=P}	0.3791134	0.8115858	1.066738
4	{NRC_MEDIUM=E,NRC_GENDER_CODE=B}	=>	{NRC_RESULT=P}	0.1077305	0.8037028	1.056377

Negative Association

	SCHL_TYPE==G	SCHL_TYPE=!G
URBAN_RURAL ==U	0.10	0.32
URBAN_RURAL !=U	0.30	0.26

	URBAN	~URBAN
PASS	0.32	0.45
~PASS	0.11	0.12

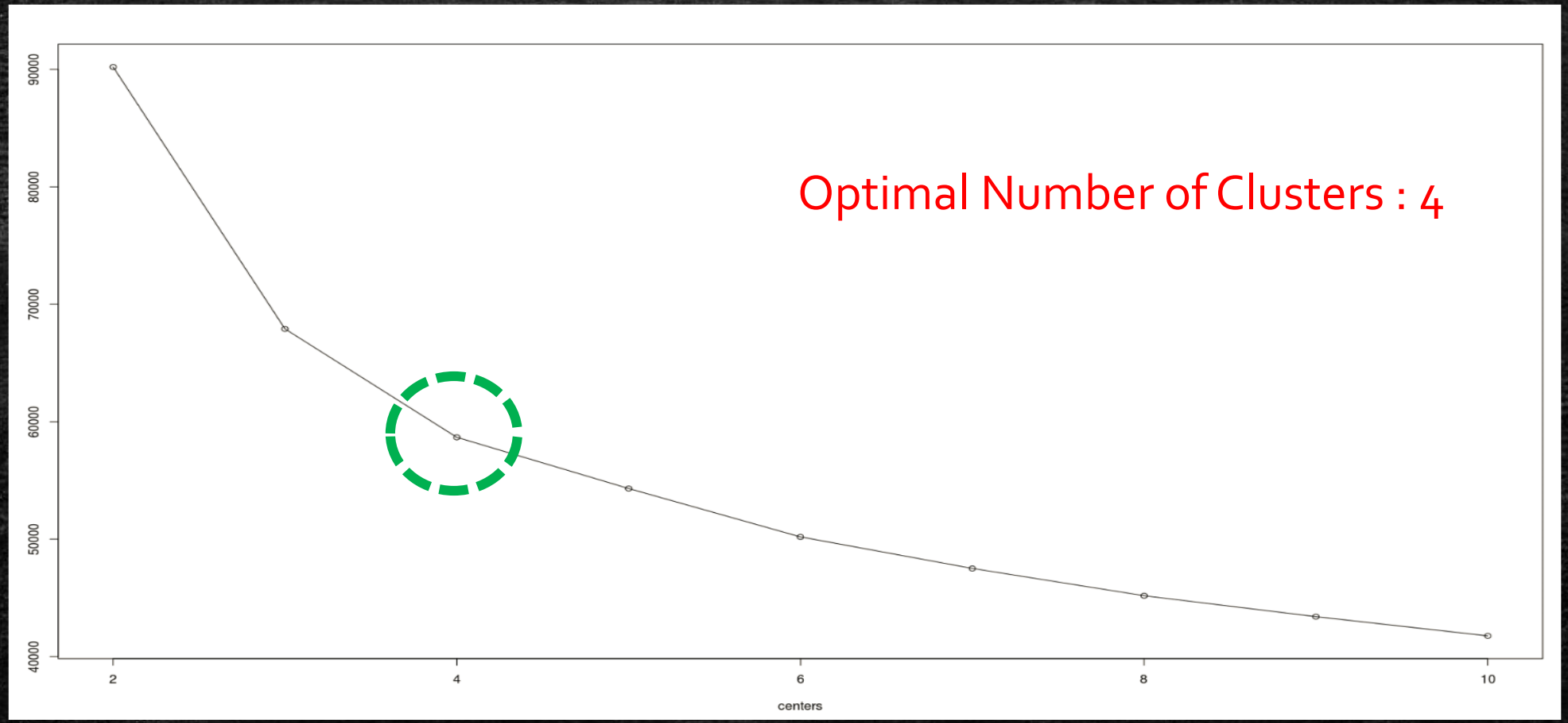
	SCHL_TYPE==U	SCHL_TYPE=!U
URBAN_RURAL ==R	0.096	0.47
URBAN_RURAL !=R	0.181	0.24

	DISTINCTION	~DISTINCTION
RURAL	0.03	0.54
~RURAL	0.05	0.38

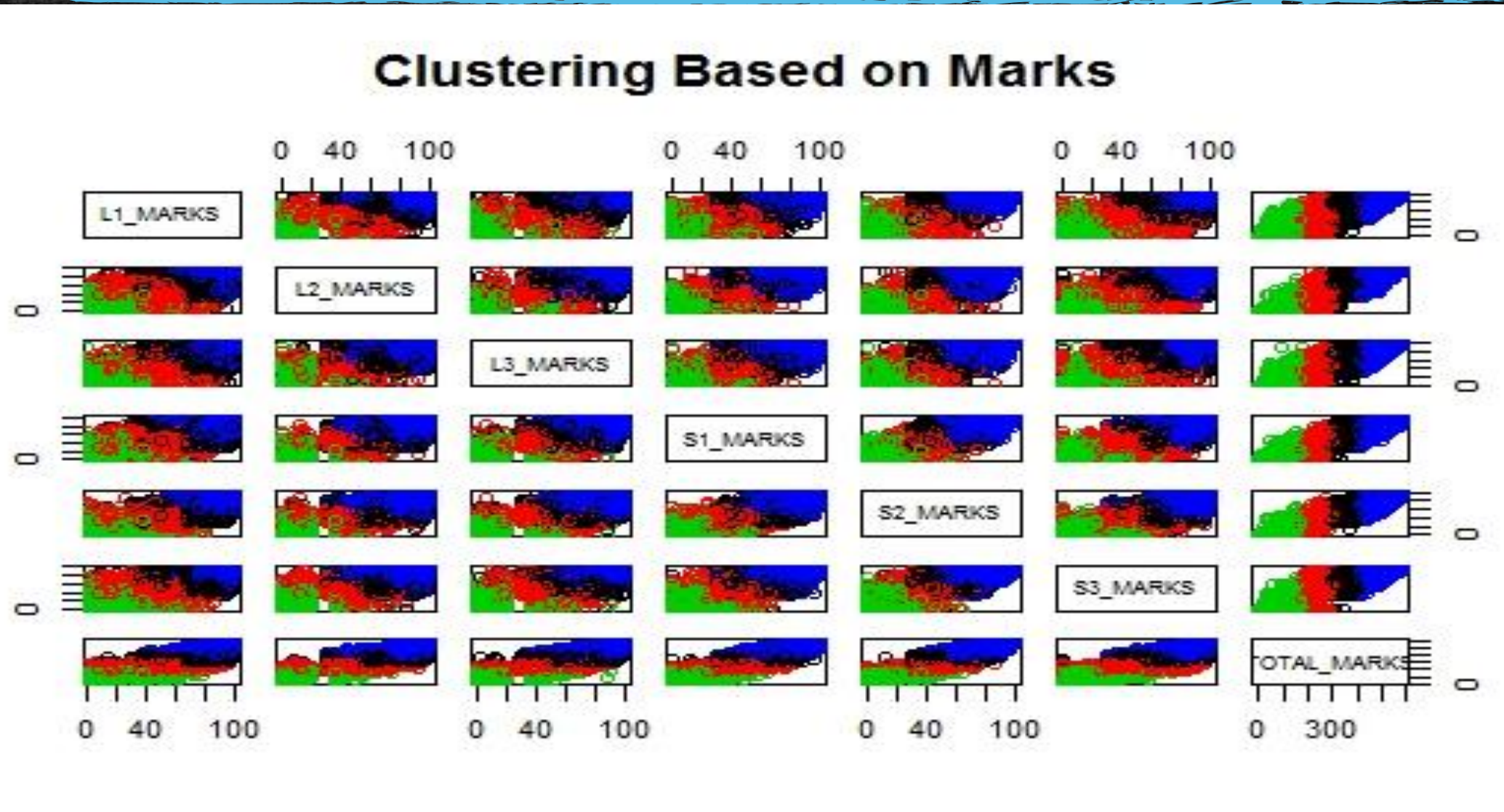
Infrequent Pattern

NRC_MEDIUM = E	L1_CODE=T
PHYSICAL_CONDITION=B	NRC_MEDIUM=E
NRC_CLASS=D	NRC_CASTE_CODE = 2

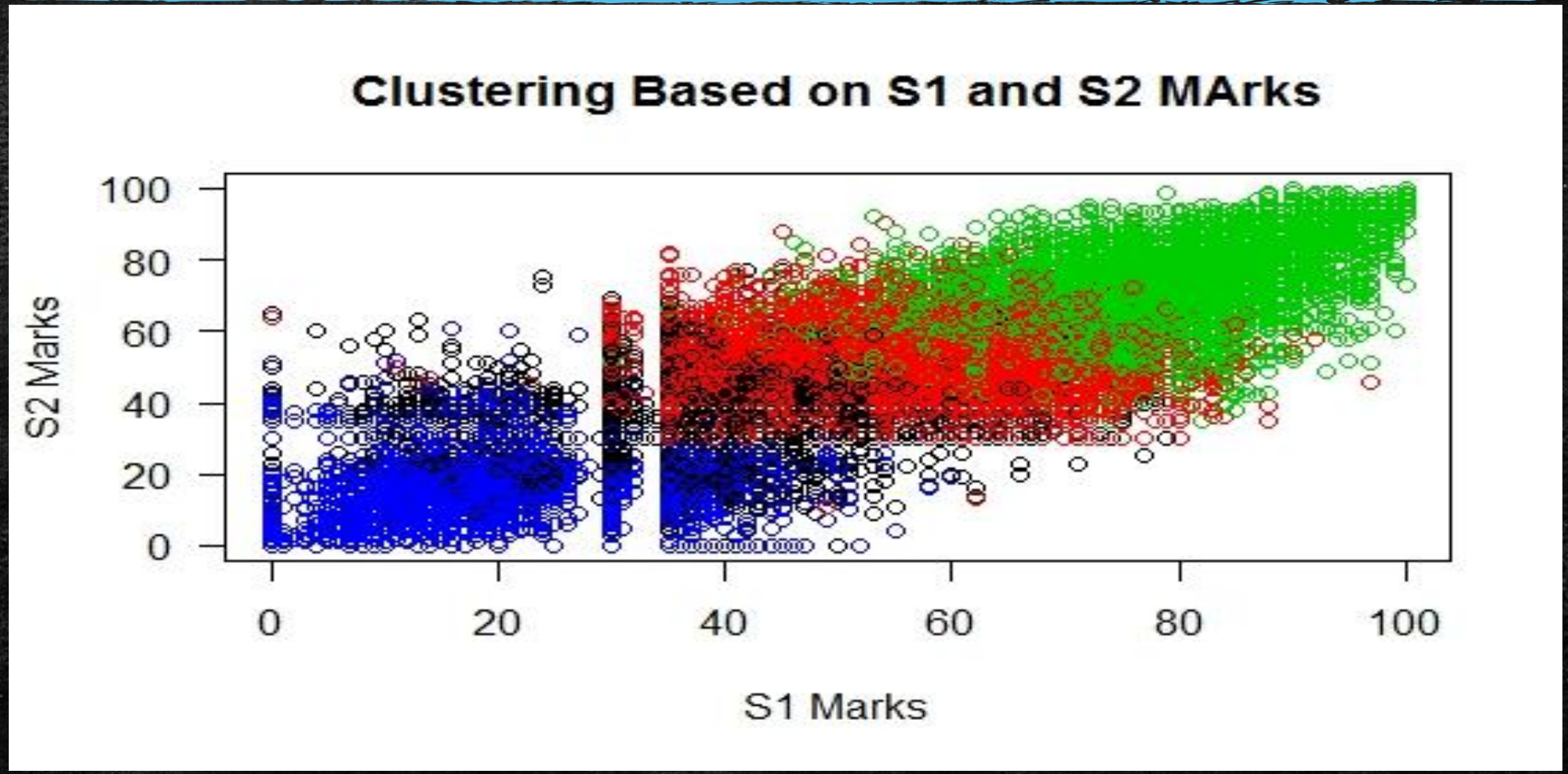
Clustering


90000
80000
70000
60000
50000
40000
2
4
6
8
10
centers
Optimal Number of Clusters : 4

K-MEANS Clustering



K MEANS


Clustering Based on S1 and S2 MArks
S2 Marks
100
80
60
40
20
0
0
20
40
60
80
100
S1 Marks

Cross Cluster Analysis

Cluster 1

Gender	CASTE CODE:	School Type
• Boy : 49.07% • Girl : 50.93%	• 1 : 19.67% • 2 : 8.78% • 3 : 8.9% • 4 : 62.54%	• A : 32.09% • G : 27.37% • U : 40.52%

Cluster 2

Gender	CASTE CODE:	School Type
• Boy : 49.65% • Girl : 50.35%	• 1 : 27.91 % • 2 : 17.99% • 3 : 14.35% • 4 : 39.45%	• A : 33.19% • G : 34.95% • U : 31.85%

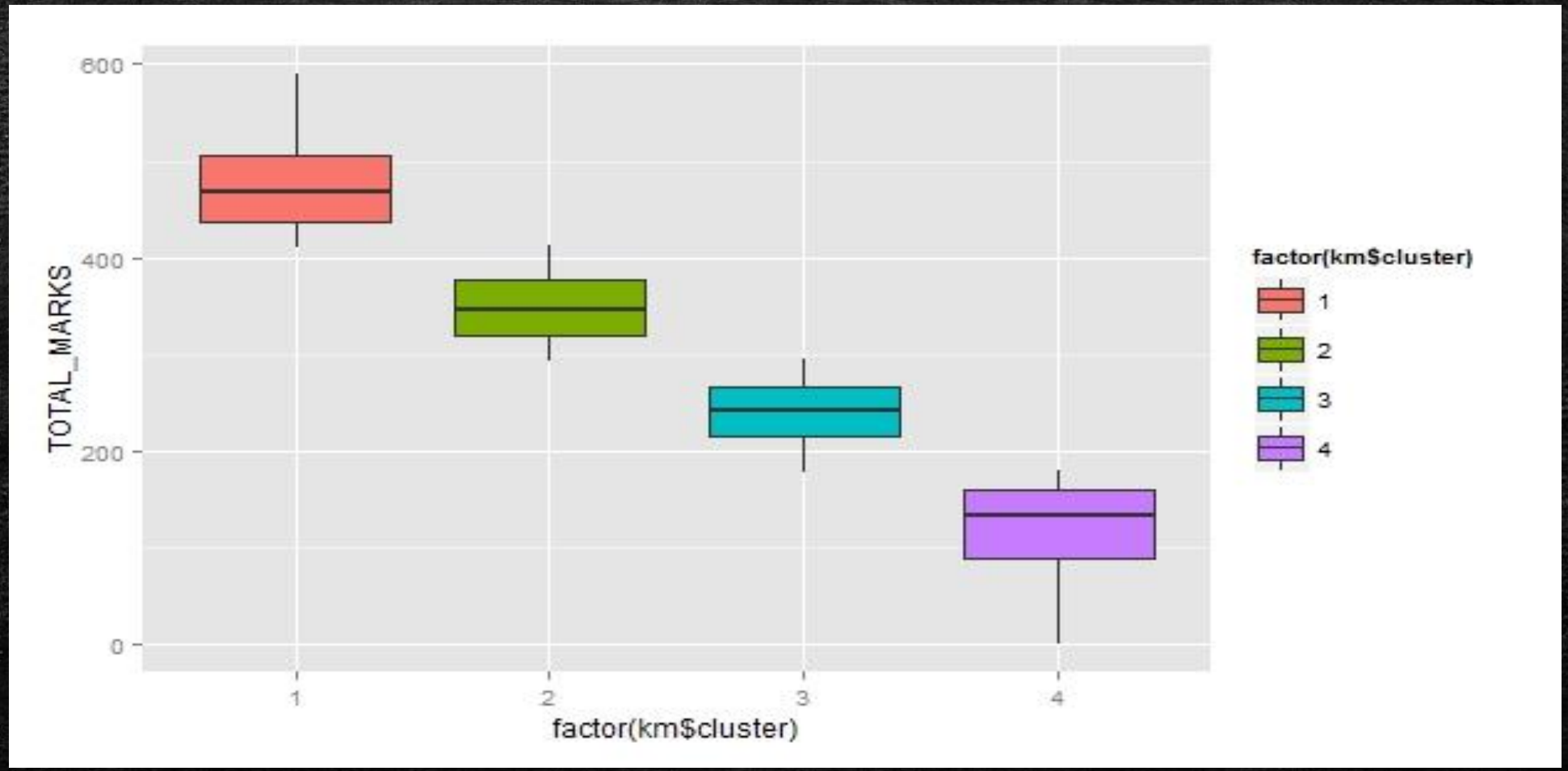
Cluster 3

Gender	CASTE CODE:	School Type
• Boy : 53.6% • Girl : 46.4%	• 1 : 27.46 % • 2 : 16.61% • 3 : 13.98% • 4 : 41.94%	• A : 33.4% • G : 39.23% • U : 27.31%

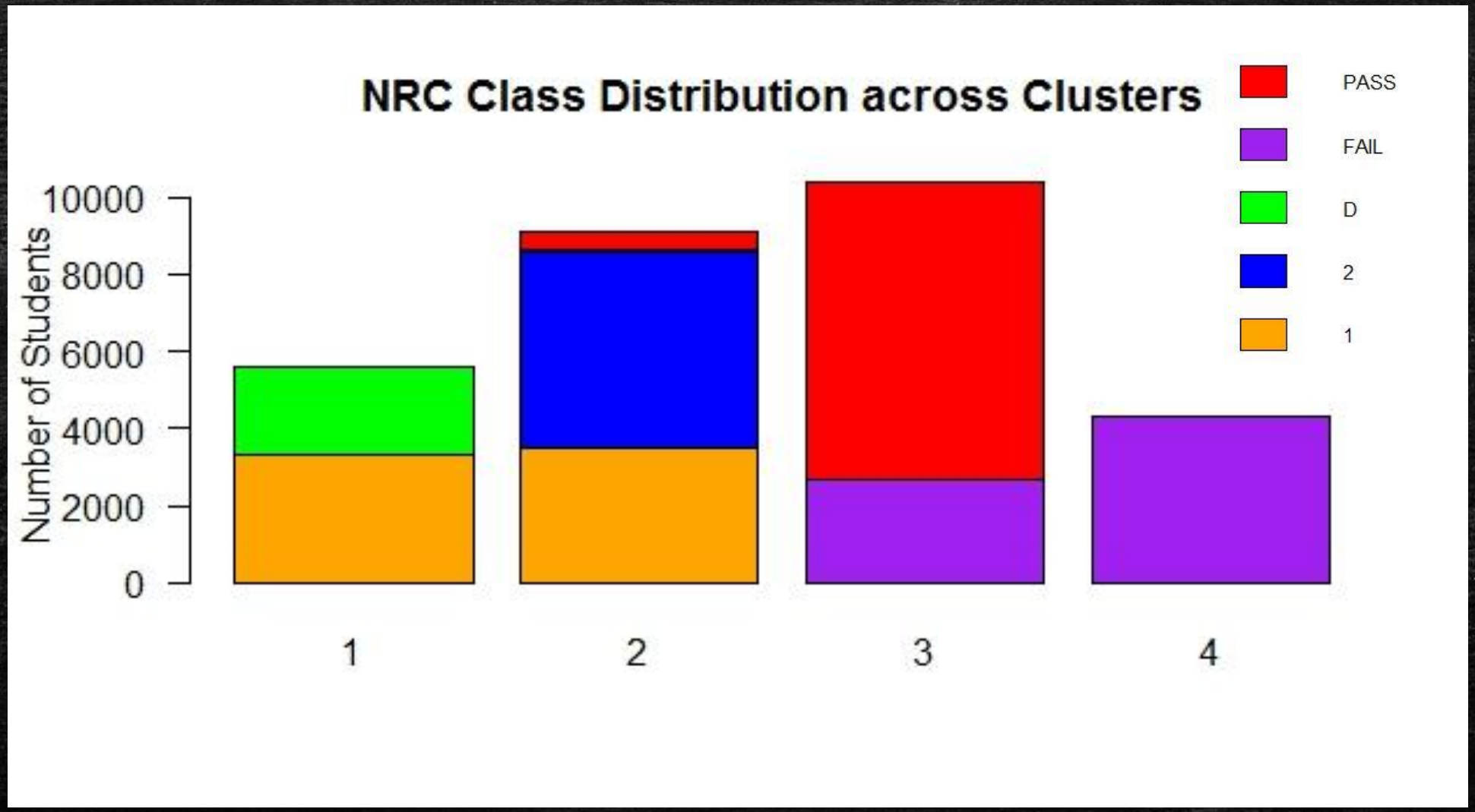
Cluster 4

Gender	CASTE CODE:	School Type
• Boy : 57.69% • Girl : 42.30%	• 1 : 31.85 % • 2 : 12.9% • 3 : 9.05% • 4 : 46.17%	• A : 30% • G : 41% • U : 29%

Cross Cluster Analysis



Cross Cluster Analysis



Have You Got A Minute?

A MINUTE OF PAIN
IS WORTH A LIFETIME
OF GLORY

-Pete Zamperini

NA Rows Analysis

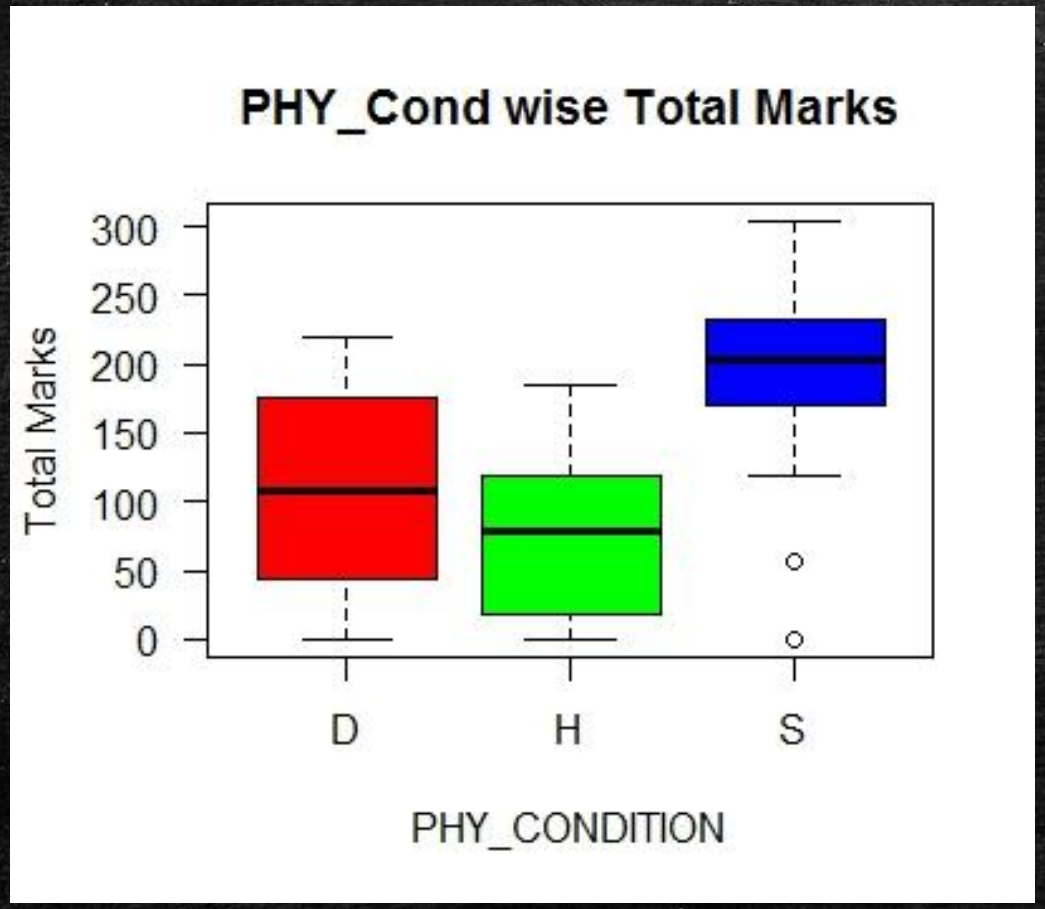
- Physical Condition
  - Dumb : 23
  - Hearing impaired : 11
  - Spastic / dyslexia : 13
- There are 15 other Schools that are not present in clean data.
- There is no L3. All L3 fields are NA. Majority of L2 also NA
  - PHY_COND candidate might have different component.
- All have registered for S1-S2-S3 subject. L1 majority is enrolled.

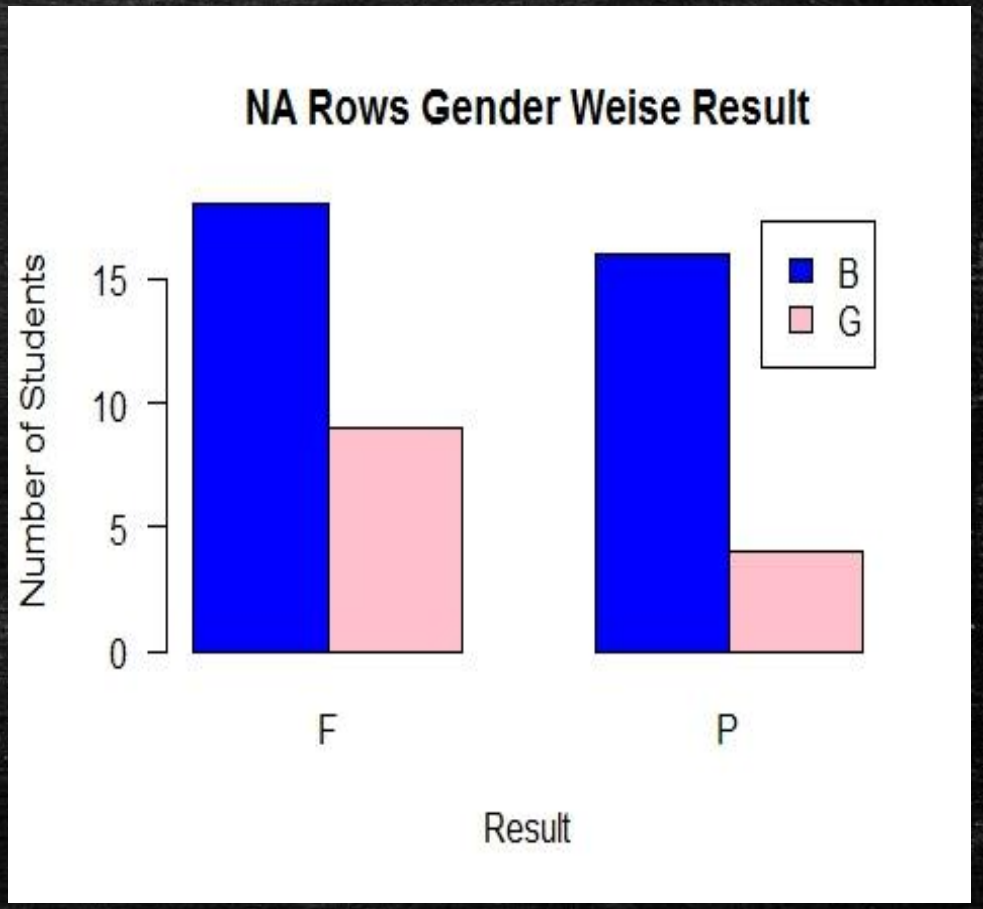
NA Rows Analysis

- School Type : A – 18 , G – 3 , U – 26
  - Majority of PHY_COND candidate prefer Unaided School. Or Extremely less number of schools for PHY candidates.
- Association Rule Mining

lhs		rhs	support	confidence	lift
{NRC_CASTE_CODE=4,NRC_MEDIUM=E}	=>	{NRC_PHYSICAL_CONDITION=S}	0.2553191	0.8571429	3.098901
{NRC_CASTE_CODE=4,NRC_PHYSICAL_CONDITION=D}	=>	{NRC_MEDIUM=K}	0.2978723	1.0000000	1.516129
{NRC_PHYSICAL_CONDITION=D}	=>	{NRC_MEDIUM=K}	0.4680851	0.9565217	1.450210
{NRC_CASTE_CODE=1,NRC_MEDIUM=K}	=>	{NRC_PHYSICAL_CONDITION=D}	0.1063830	1.0000000	2.043478

NA Rows Analysis

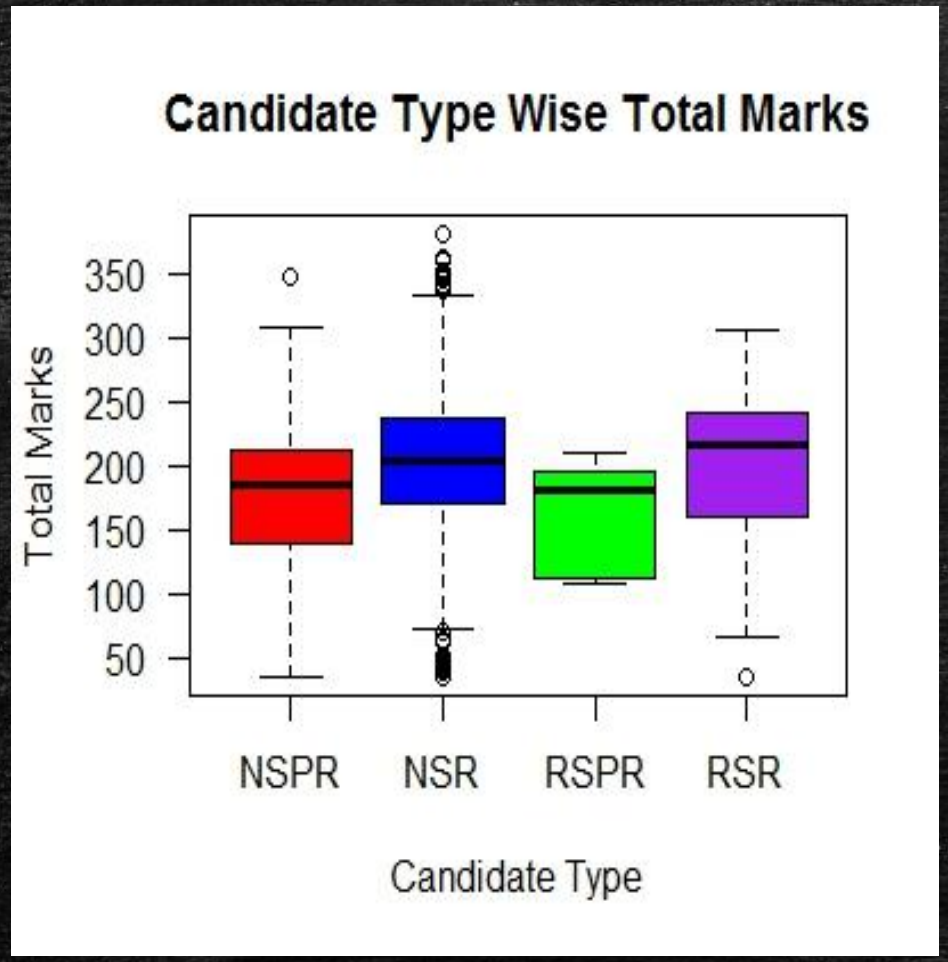




Star Value Rows Analysis

- These record represent NSPR, NSR, RSPR, RSR type of Candidates.
- 3 Schools are there that are not present in Cleaned Data.
- All the candidate falling in this category has NRC_CLASS as PASS or FAIL. No one has got 1/2/D Class.

Star Value Rows Analysis



Nupur Garg

Setu Patani

Praveen Baby