SUPERVISED LEARNING
FINAL GRADE REGRESSION

By Trent Casillas

DATA SET DESCRIPTION

The data set used looks over study, school habits, gender , personal , and family history. There is a mix of 33 categorical and numeric variables. Dataset can be found at https://www.kaggle.com/uciml/student-alcohol-consumption/home.

school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

sex - student's sex (binary: 'F' - female or 'M' - male)

age - student's age (numeric: from 15 to 22)

address - student's home address type (binary: 'U' - urban or 'R' - rural)

famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

guardian - student's guardian (nominal: 'mother', 'father' or 'other')

traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

failures - number of past class failures (numeric: n if 1<=n<3, else 4)

schoolsup - extra educational support (binary: yes or no)

famsup - family educational support (binary: yes or no)

paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

activities - extra-curricular activities (binary: yes or no)

nursery - attended nursery school (binary: yes or no)

higher - wants to take higher education (binary: yes or no)

internet - Internet access at home (binary: yes or no)

romantic - with a romantic relationship (binary: yes or no)

famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

freetime - free time after school (numeric: from 1 - very low to 5 - very high)

goout - going out with friends (numeric: from 1 - very low to 5 - very high)

Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

health - current health status (numeric: from 1 - very bad to 5 - very good)

absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

G1 - first period grade (numeric: from 0 to 20)

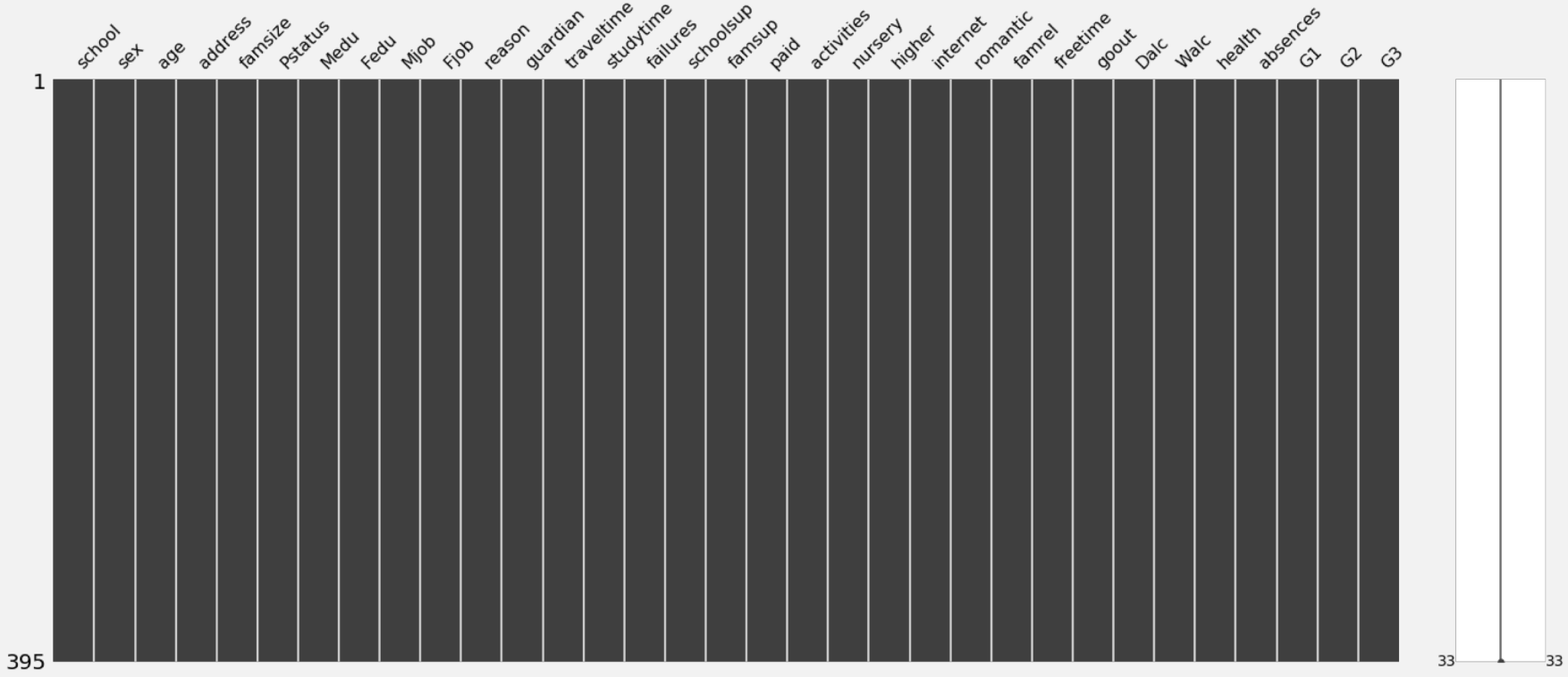
G2 - second period grade (numeric: from 0 to 20)

G3 - final grade (numeric: from 0 to 20, output target)

RESEARCH QUESTION

- This presentation will predict the final grade(G3) via regression methods: KNN, Lasso, Ridge, and Support Vector Machine Regression.

DATASET REVIEW



DATA CLEANING

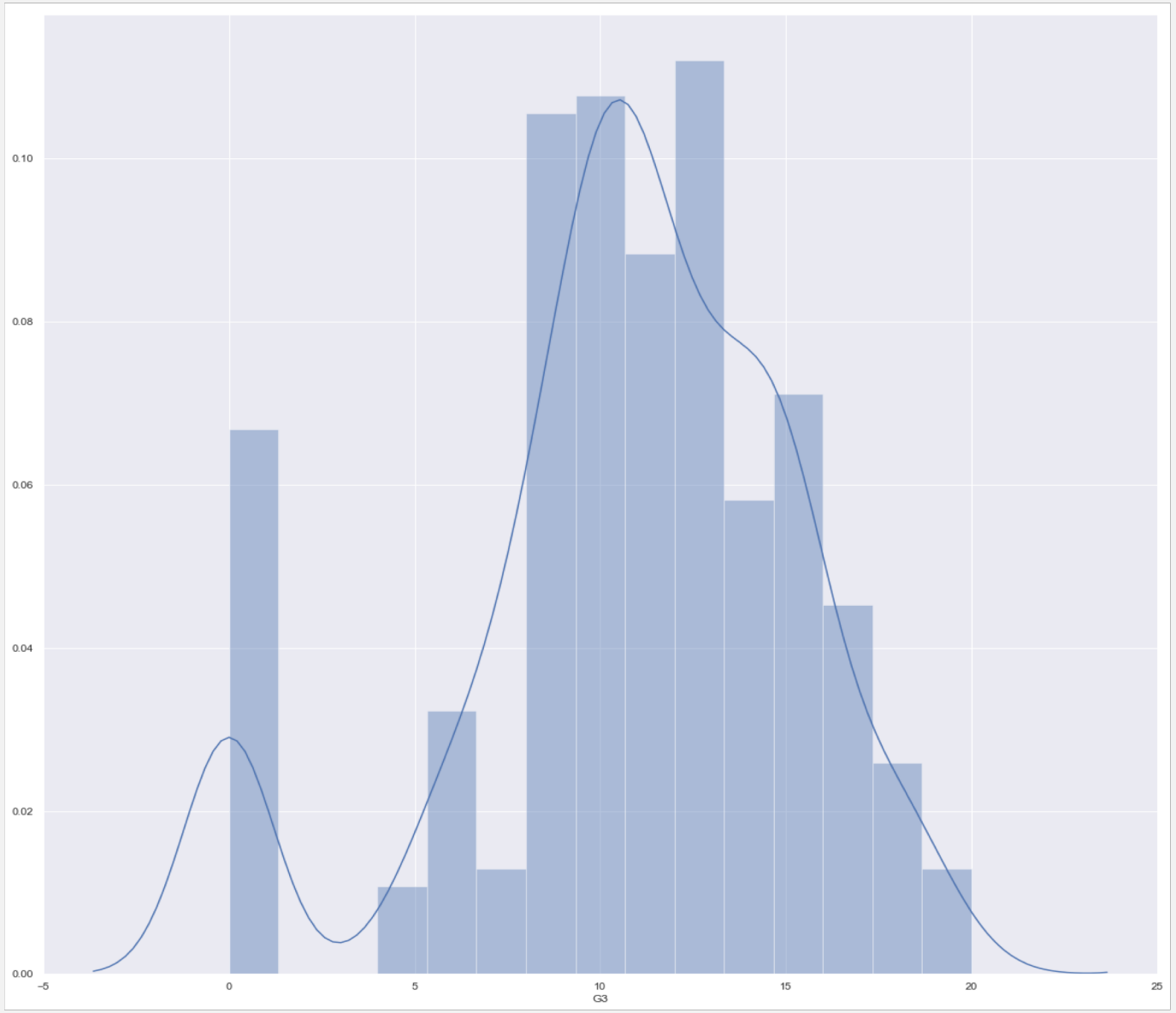
ORIGINAL

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
count	395	395	395	395	395	395	395	395	395	395	395	395	395	395	395	395
mean	16.6962	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304	3.235443	3.108861	1.481013	2.291139	3.55443	5.708861	10.90886	10.71392	10.41519
std	1.276043	1.094735	1.088201	0.697505	0.83924	0.743651	0.896659	0.998862	1.113278	0.890741	1.287897	1.390303	8.003096	3.319195	3.761505	4.581443
min	15	0	0	1	1	0	1	1	1	1	1	1	0	3	0	0
25%	16	2	2	1	1	0	4	3	2	1	1	3	0	8	9	8
50%	17	3	2	1	2	0	4	3	3	1	2	4	4	11	11	11
75%	18	4	3	2	2	0	5	4	4	2	3	5	8	13	13	14
max	22	4	4	4	4	3	5	5	5	5	5	5	75	19	19	20

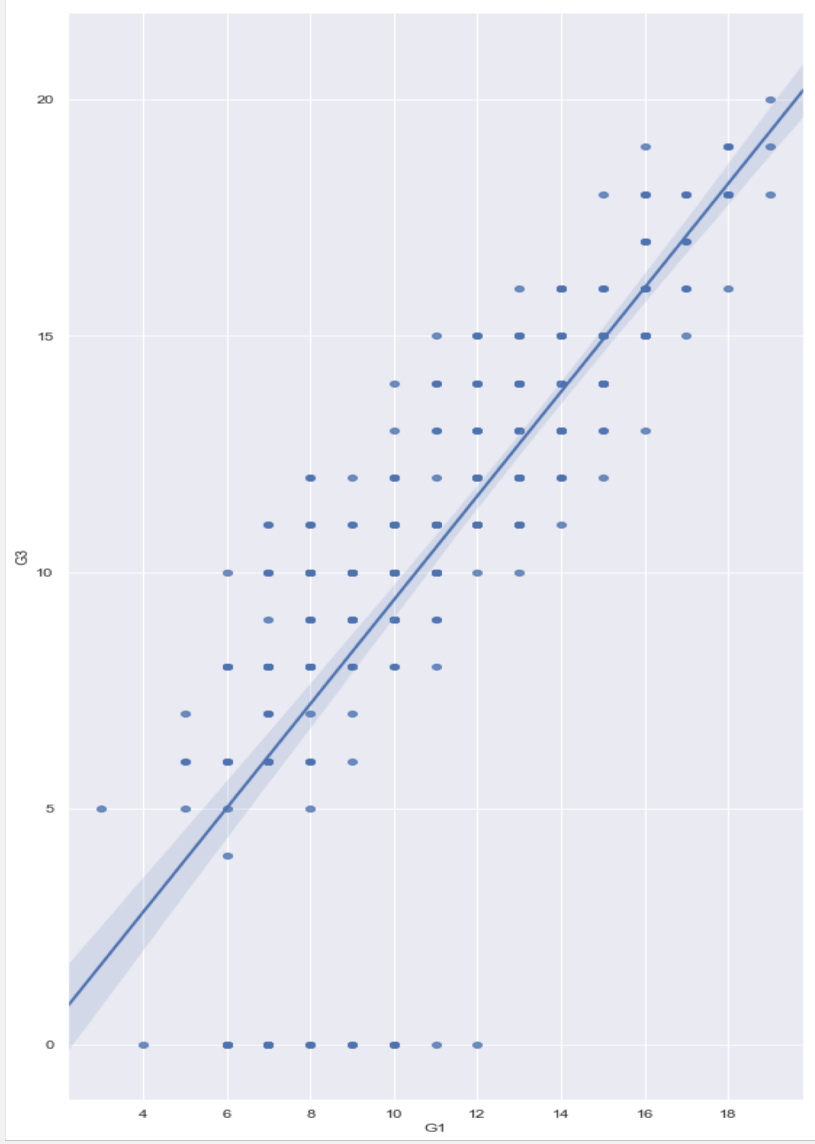
REMOVING OUTLIERS

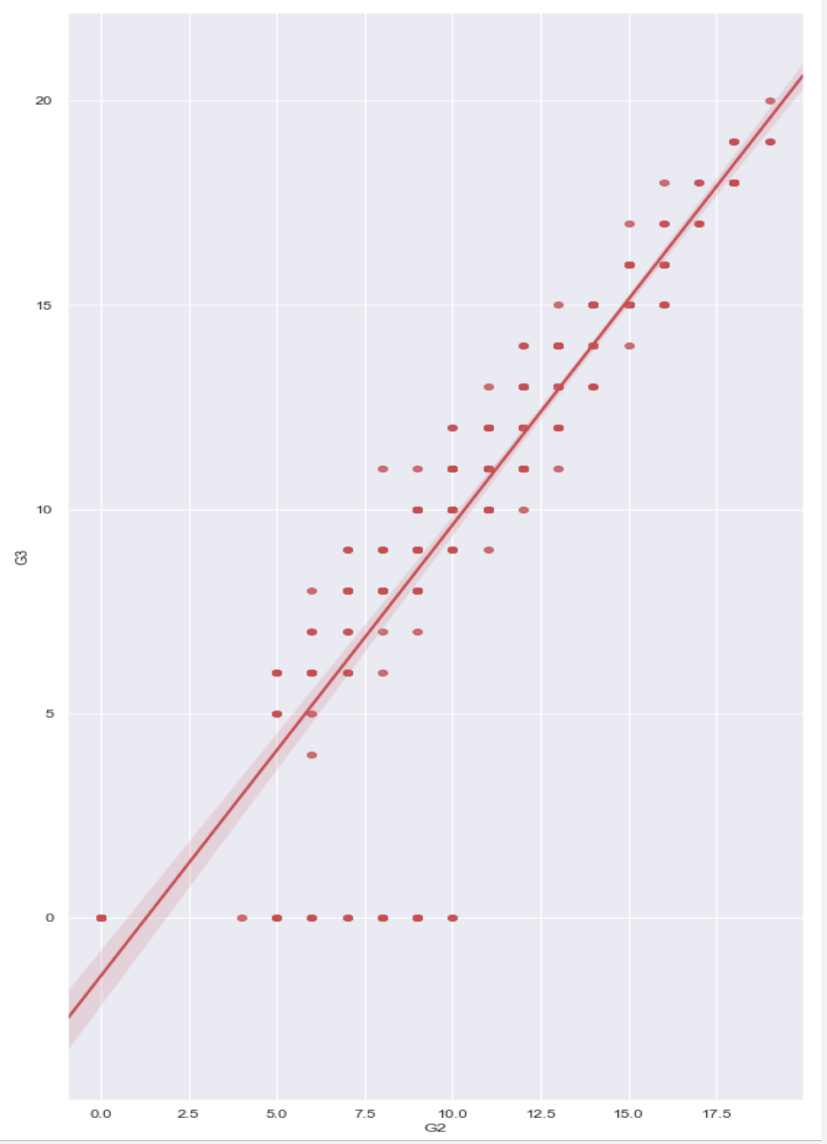
	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
count	348	348	348	348	348	348	348	348	348	348	348	348	348	348	348	348
mean	16.62931	2.784483	2.563218	1.393678	2.074713	0.212644	3.994253	3.20977	3.063218	1.324713	2.16092	3.54023	4.945402	11.1523	10.96264	10.67816
std	1.213732	1.082792	1.078489	0.595531	0.845496	0.504345	0.788952	0.977686	1.069096	0.593693	1.174892	1.39207	5.852156	3.326138	3.734918	4.564603
min	15	0	0	1	1	0	2	1	1	1	1	1	0	3	0	0
25%	16	2	2	1	1.75	0	4	3	2	1	1	3	0	8	9	9
50%	17	3	3	1	2	0	4	3	3	1	2	4	3	11	11	11
75%	18	4	3.25	2	3	0	5	4	4	2	3	5	7	14	14	14
max	20	4	4	3	4	2	5	5	5	3	5	5	30	19	19	20

TARGET DISTRIBUTION



G1 AND G2 FEATURE CORRELATION WITH G3

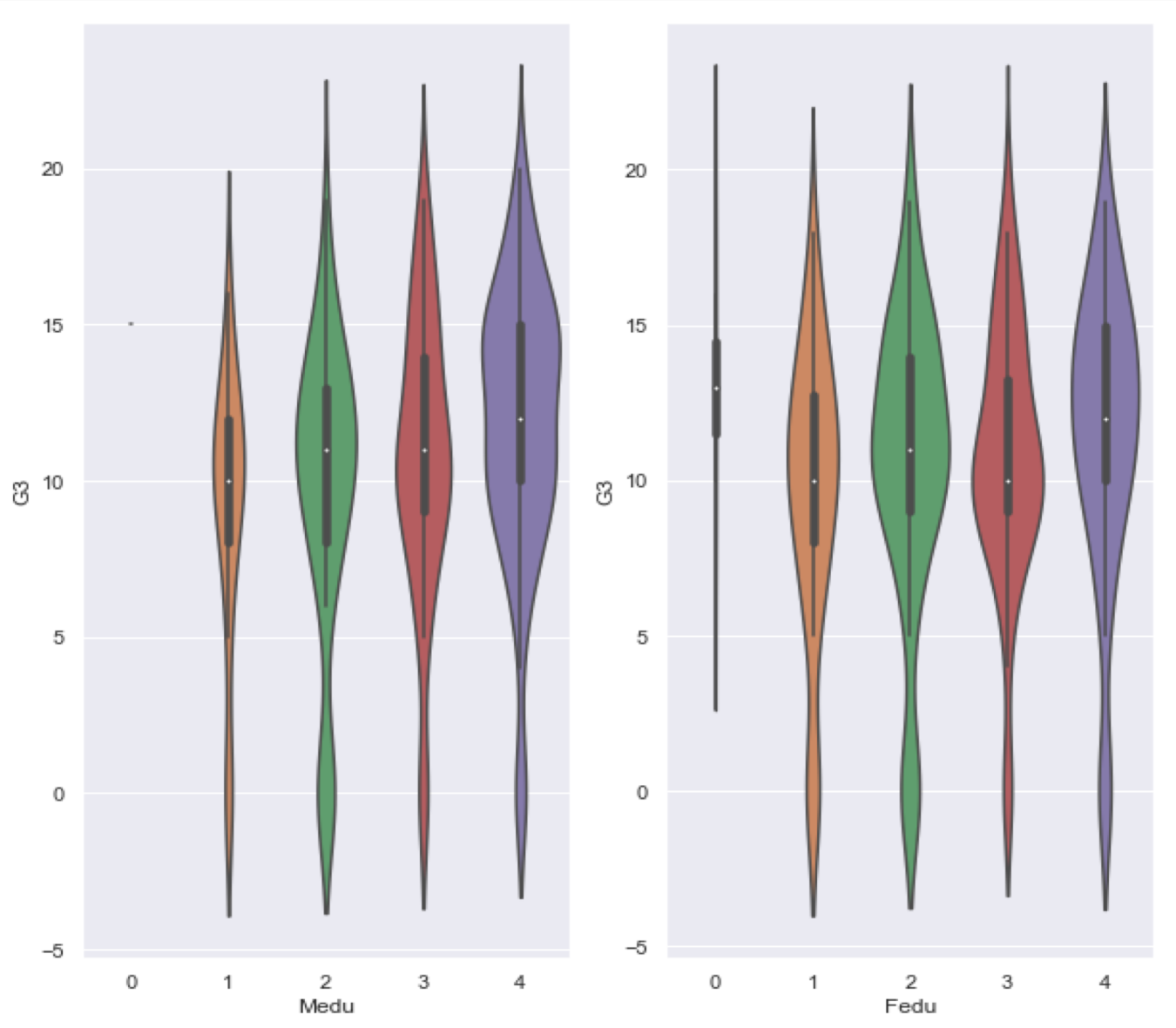




G1 - first period grade (numeric: from 0 to 20)

G2 - second period grade (numeric: from 0 to 20)

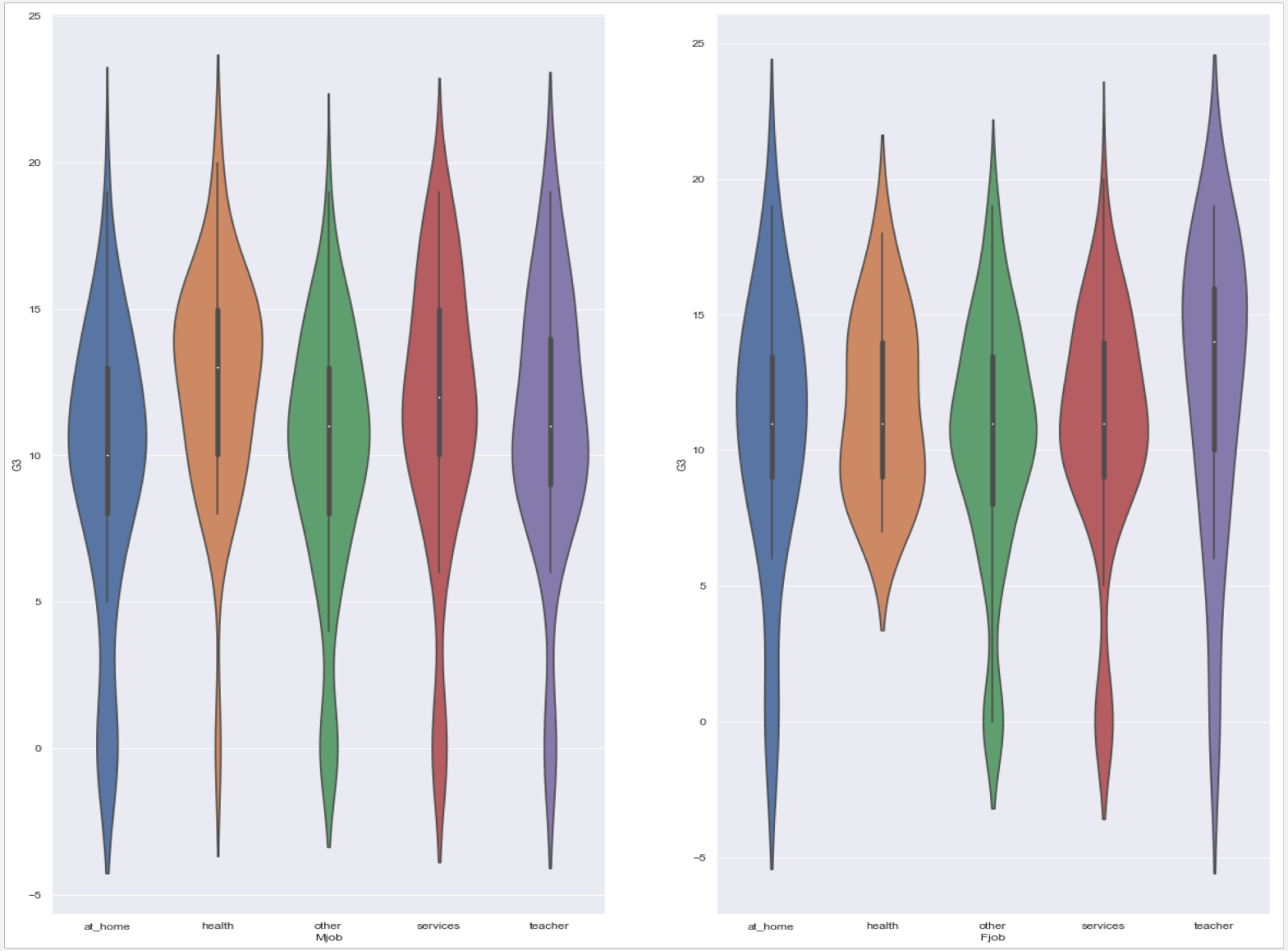
MOTHER AND FATHER EDUCATION



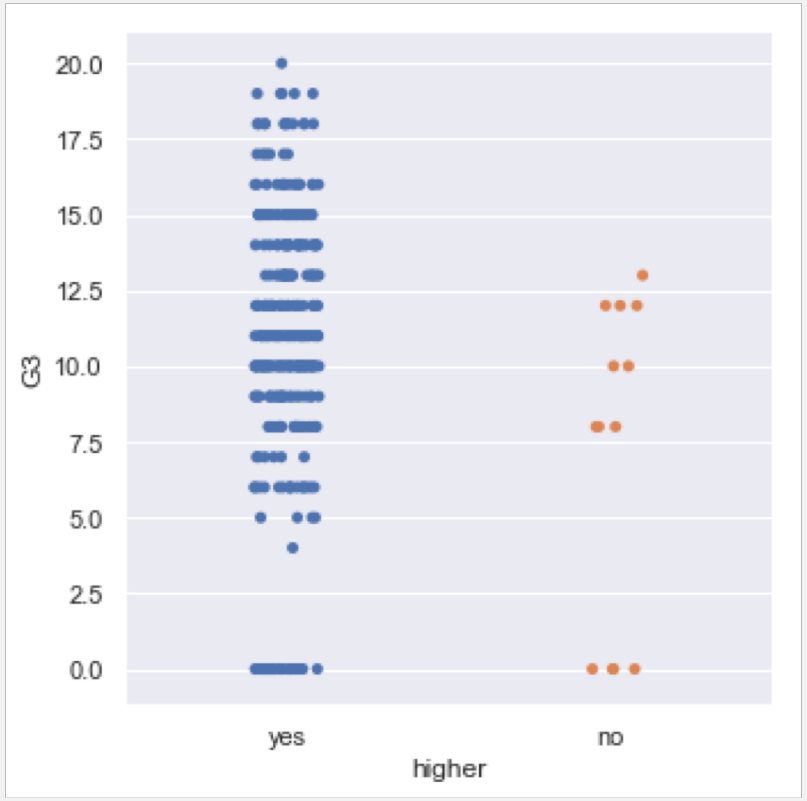
Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

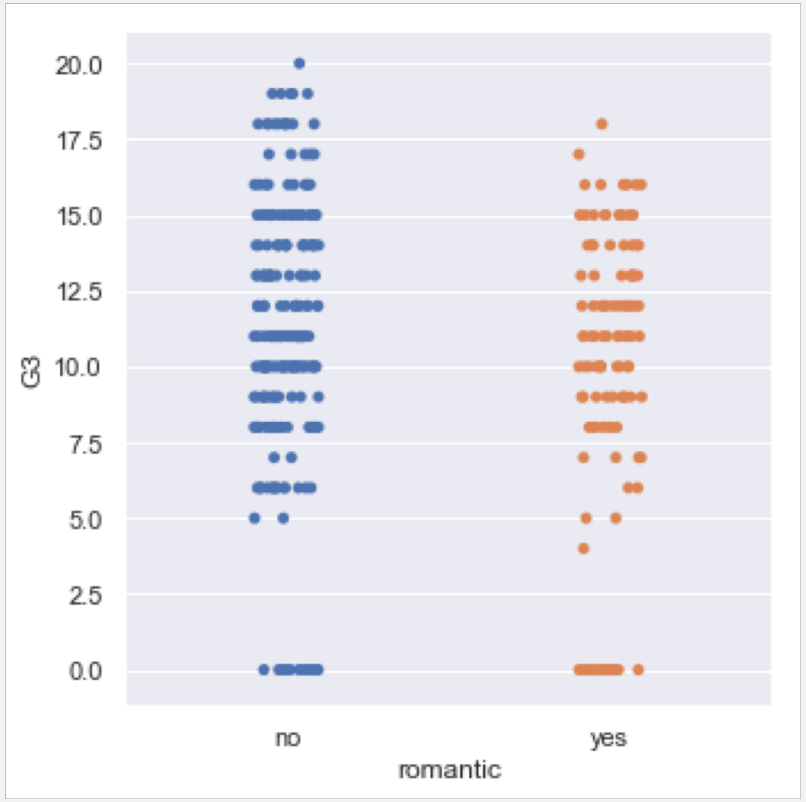
Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

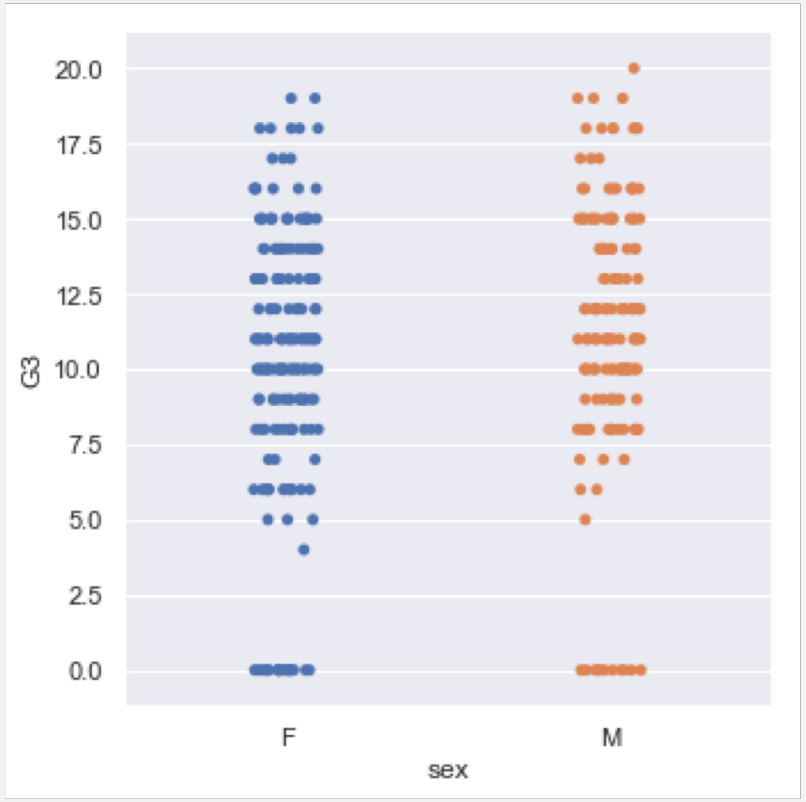
MOTHER AND FATHER'S JOBS



HIGHER EDUCATION, RELATIONSHIPS, AND GENDER







FEATURE AND MODEL SELECTION

Top 10 Positive Correlated Features	
G2	0.904496
G1	0.802543
Medu	0.20108
higher_yes	0.152316
romantic_no	0.144539
sex_M	0.138167
Mjob_health	0.130446
Fedu	0.124737
address_U	0.119381
schoolsup_no	0.104309
Fjob_teacher	0.102654

Top 10 Negative Correlated Features	
failures	-0.319443
age	-0.15684
higher_no	-0.152316
romantic_yes	-0.144539
sex_F	-0.138167
traveltime	-0.128179
goout	-0.121014
address_R	-0.119381
Mjob_other	-0.11124
Mjob_at_home	-0.108338
schoolsup_yes	-0.104309

- 59 Total features after applying get dummies.
- Feature selection was done by using a portion of both ends combining the negative and positive correlated features.
- Ultimately, all the features were used.
- Negative values were taken as 0 and values over 20 were taken as 20.

Training Model Selection:
X=df.drop('G3',1)
Y=df.G3

- Lasso, Ridge, KNN, and SVM regression were looked at for modeling.

LASSO

Parameters:

- alpha=0.25 & selection='random'
- When alpha increased to 15, $R^2 \rightarrow 0$.
- R^2: 0.8319070479732708
- Cross Validation Scores [0.87875054 0.91876338 0.83269706 0.7684241 0.71403356]. 0.82 (+/- 0.15)

	Y	Y_pred	Y_res
count	348	348	348
mean	10.67816	10.70368	-6.9E-16
std	4.564603	4.014612	1.871448
min	0	0	-9.12043
25%	9	8.196481	-0.3284
50%	11	10.54341	0.309254
75%	14	13.65803	0.948518
max	20	20	3.824633

RIDGE

Parameters:

- alpha=0.25 , fit_intercept=True, solver='auto',random_state=65
- alpha was much more consistent, solver worked consistently
- R^2: 0.8533859256970854
- Cross Validation Scores [0.83675733 0.86297155 0.76231289 0.74747365 0.69618771] 0.78 (+/- 0.12)

	Y	Y_pred	Y_res
count	348	348	348
mean	10.67816	10.70321	-1.7E-15
std	4.564603	4.137561	1.747796
min	0	0	-7.54739
25%	9	8.089433	-0.61139
50%	11	10.66207	0.227189
75%	14	13.51651	0.984081
max	20	20	4.888826

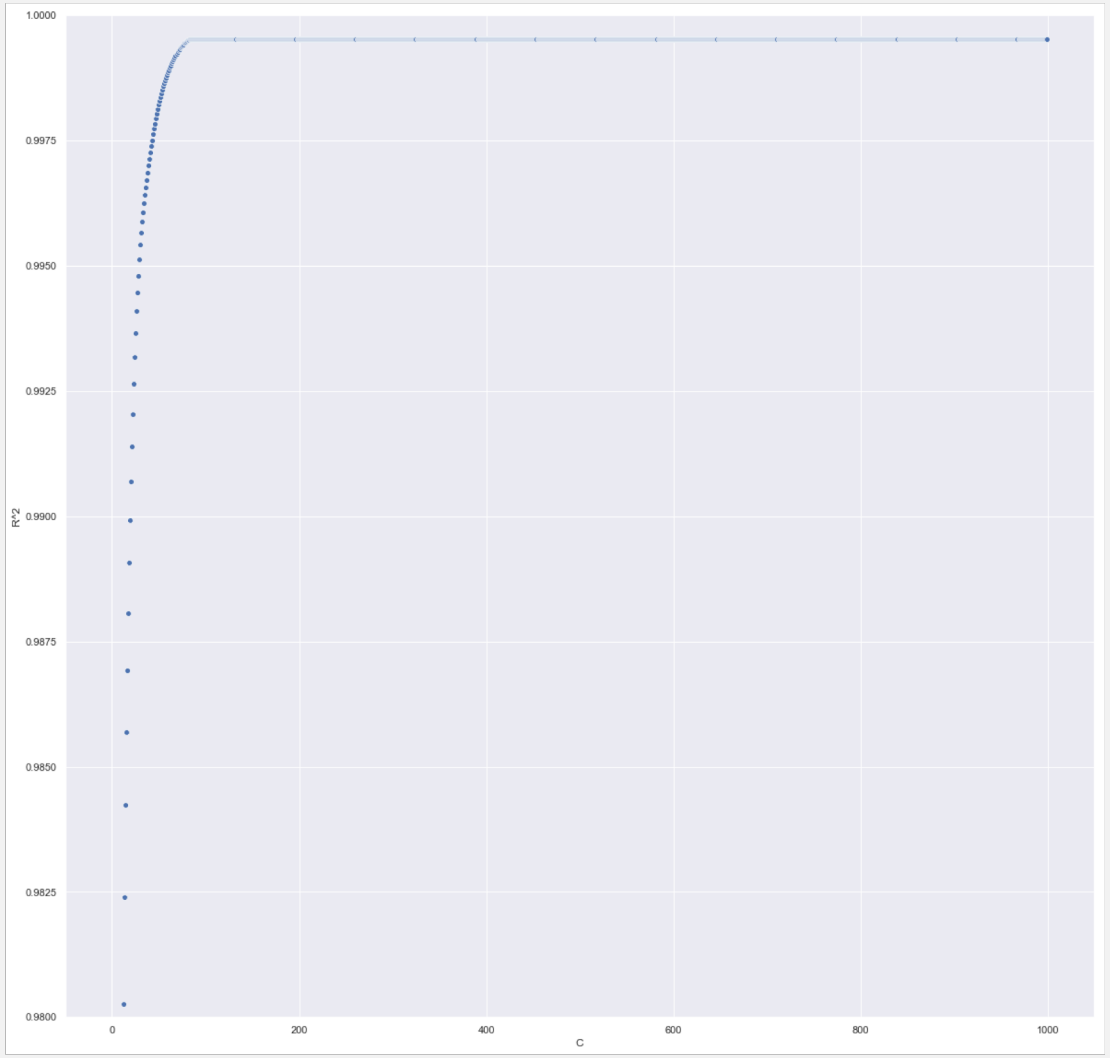
SVR

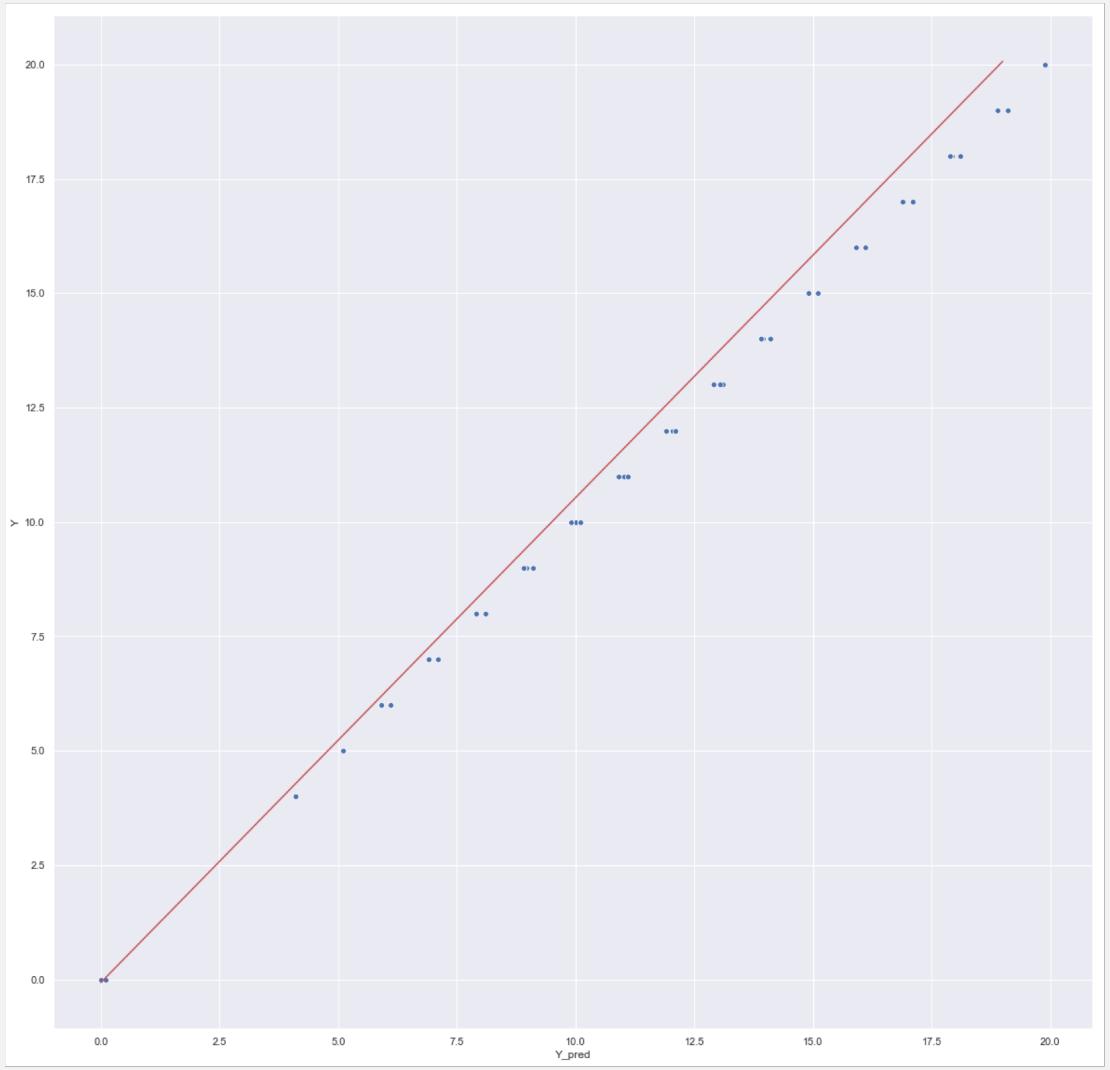
Parameters:

- Gamma='auto'
- R^2: 0.8245801277332412
- Cross Validation Scores [0.8800475 0.72897754 0.73949841 0.67918336 0.6354082] 0.73 (+/- 0.17)

	Y	Y_pred	Y_res
count	348	348	348
mean	10.67816	11.01222	-0.33406
std	4.564603	3.374667	1.882301
min	0	2.264643	-8.9039
25%	9	8.778754	-0.39819
50%	11	10.90362	-0.00254
75%	14	13.717	0.414709
max	20	18.10007	5.196925

SVR (C INCREASING EFFECT)



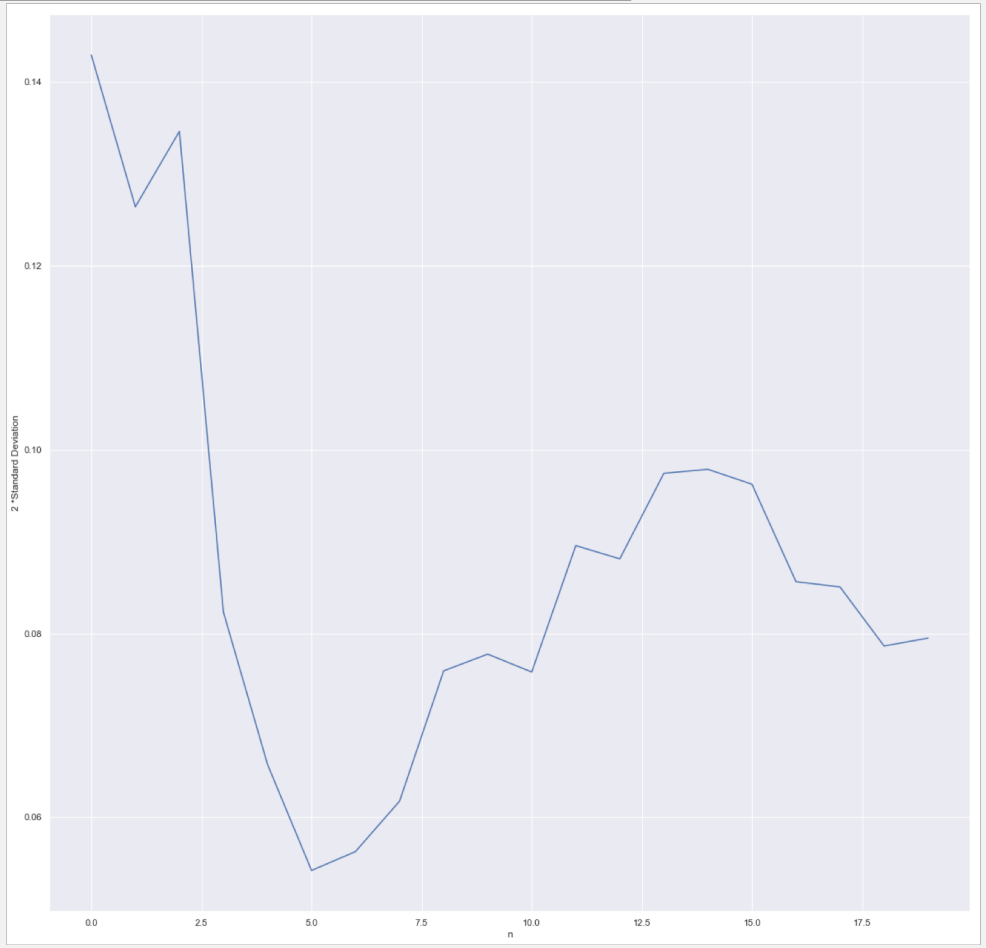


KNN REGRESSION

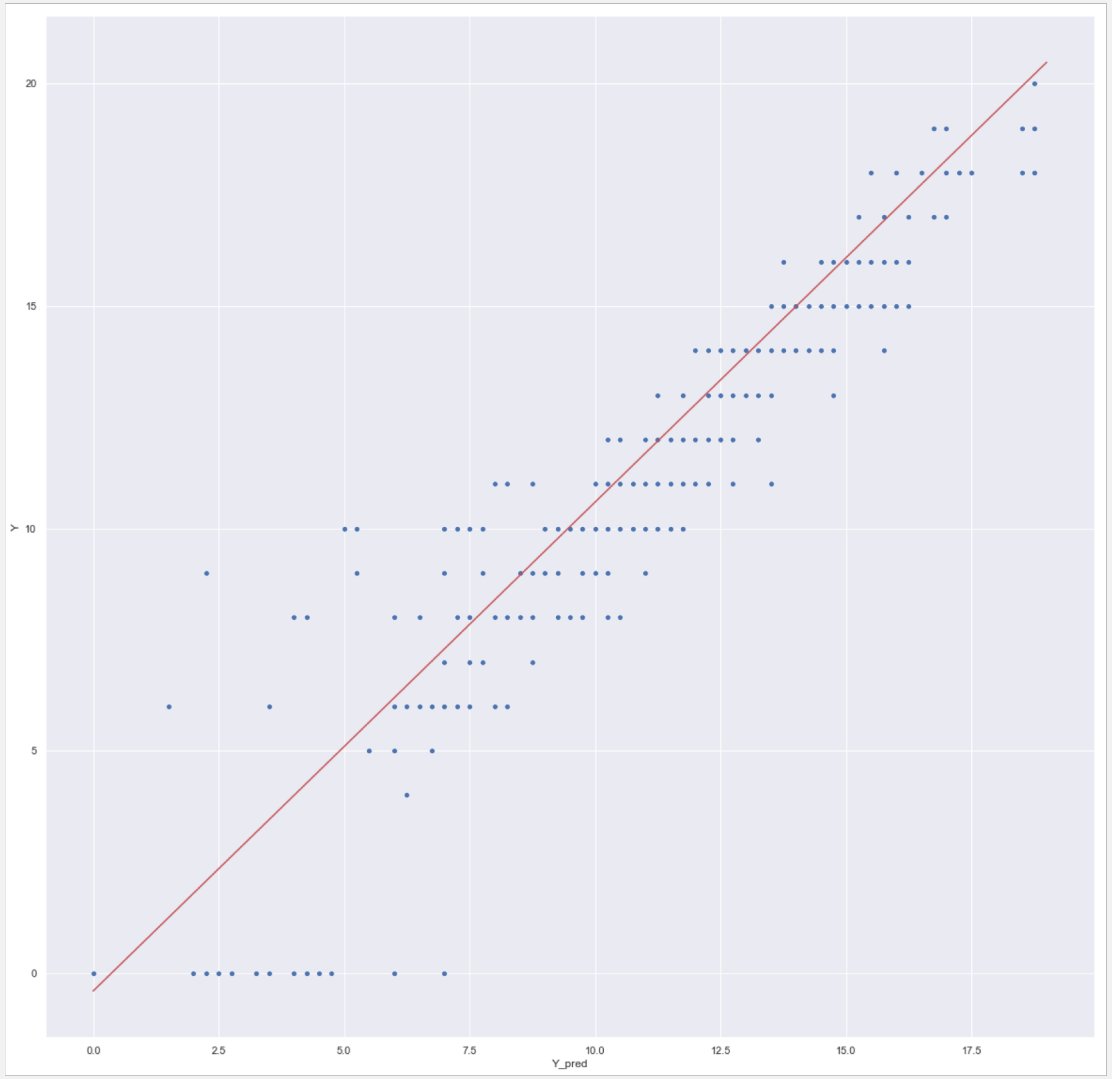
Parameters:

- n=4
- R^2: 0.899964785709516 3
- Cross Validation Scores [0.76197198 0.7730097 0.74278655 0.86001605 0.80644515] 0.79 (+/- 0.08)

	Y	Y_pred	Y_res
count	348	348	348
mean	10.67816	10.61207	0.066092
std	4.564603	4.150754	1.44219
min	0	0	-7
25%	9	8.5	-0.5
50%	11	11	0
75%	14	13.5	0.75
max	20	18.75	6.75



REGRESSION TESTING MODEL SELECTION

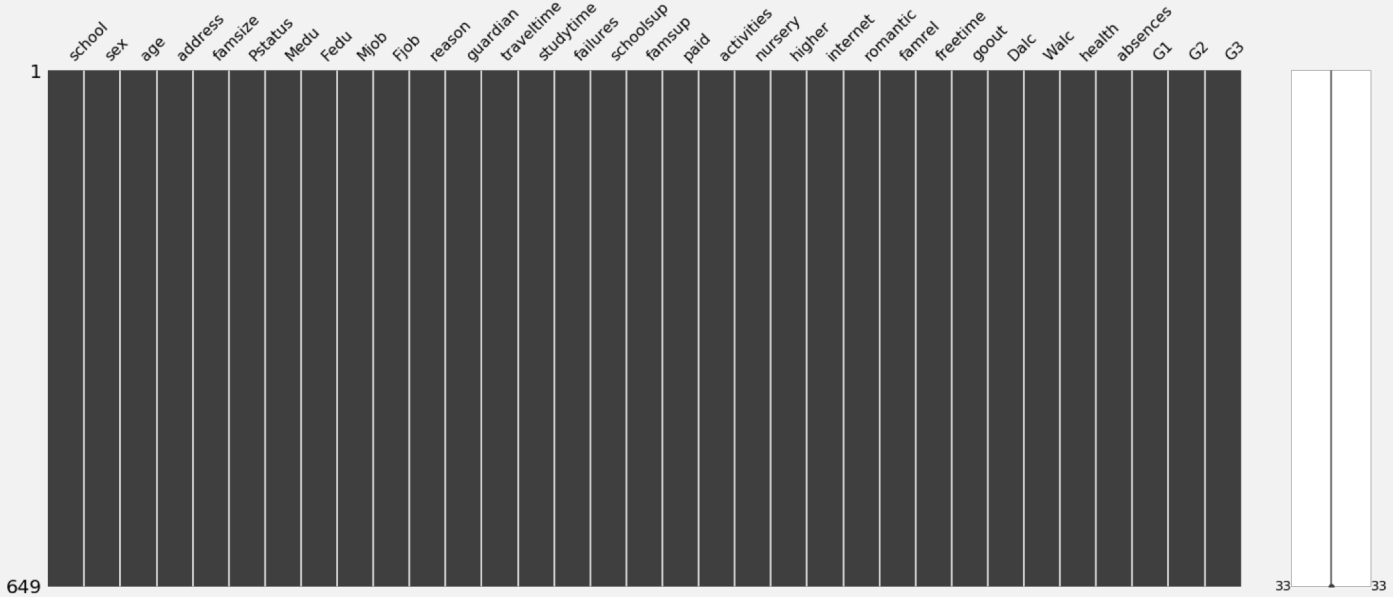


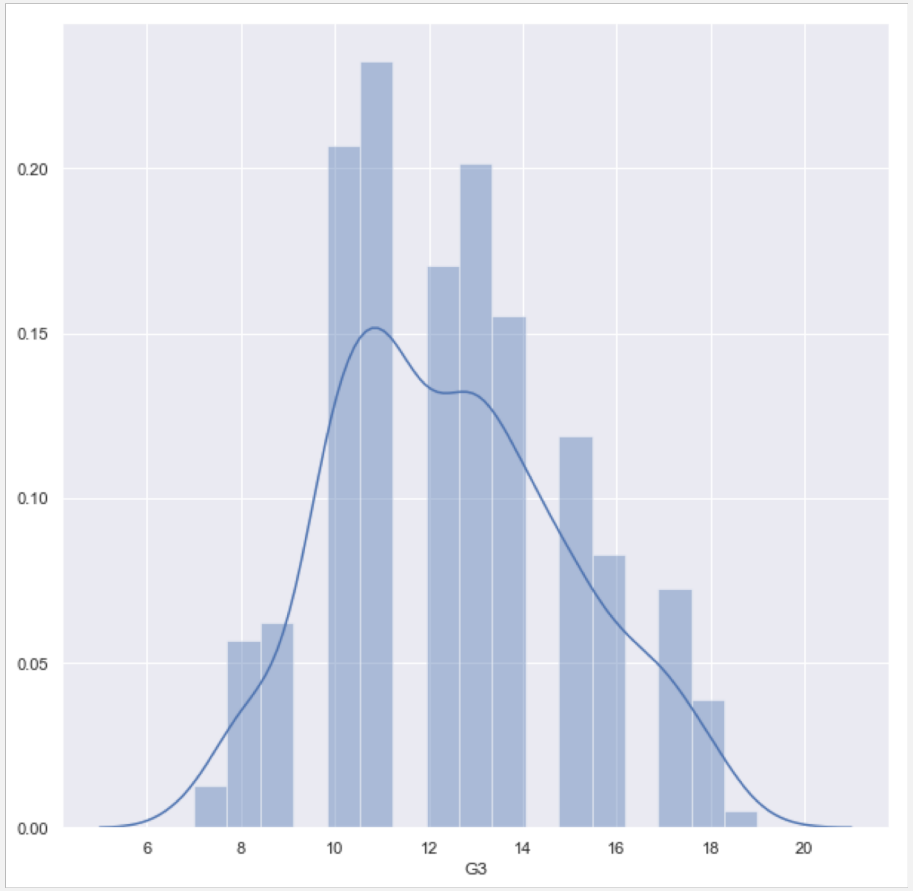
	R^2 Scores by Method:	CV.Avg ,(+/- 2* CV.Std)
Lasso:	0.8319070479732708	0.82, (+/- 0.15)
Ridge:	0.8533850845642538	0.78 (+/- 0.12)
KNN:	0.899964785709516 3	0.79, (+/- 0.08)
SVR:	0.8245801277332412	0.73, (+/- 0.17)

KNN model chosen due:

- R^2 values
- The lowest standard deviation for the cross validation.
- Easy to change around inputs in comparison to other models.

TESTING MODEL





The Portugese student data set will be used in the test model using KNN regression. The test data set is about twice as large. Outliers are removed leaving 549 rows. The distribution fits more normal as well.

FEATURE AND MODEL SELECTION

Top 10 Positive Correlated Features	
G2	0.880083
G1	0.813585
Medu	0.208806
sex_M	0.17676
Fedu	0.142634
schoolsup_no	0.138365
higher_yes	0.137908
Mjob_services	0.12467
romantic_no	0.120348
Mjob_health	0.10501

Top 10 Negative Correlated Features	
failures	-0.299473
age	-0.24164
sex_F	-0.17676
schoolsup_yes	-0.138365
higher_no	-0.137908
romantic_yes	-0.120348
Mjob_at_home	-0.117072
guardian_other	-0.10899
Mjob_other	-0.103681
internet_no	-0.097842

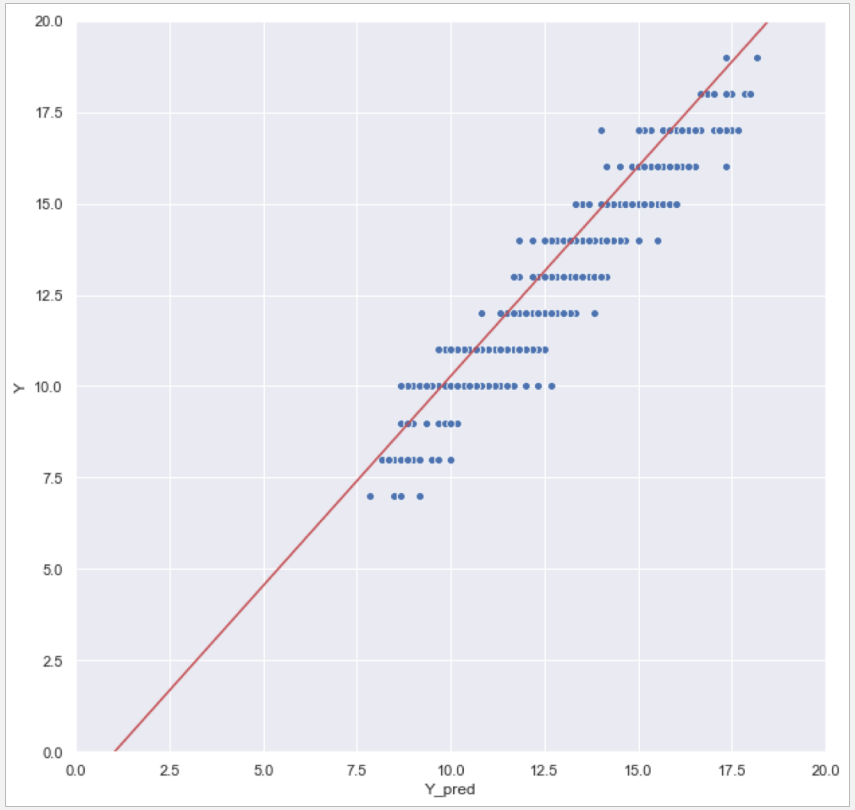
- 59 Total features after applying get dummies.
- All the features were used.
- Negative values were taken as 0 and values over 20 were taken as 20.
- Similar features are correlated with the training set.

KNN REGRESSION

Parameters:

- N=6
- R^2: 0.9003474644929953
- Cross Validation Scores [0.76211073 0.8113527 0.80079696 0.85576041 0.87652547] 0.82 (+/- 0.08)

	Y	Y_pred	Y_res
count	548	548	548
mean	12.49818	12.56691	-0.06874
std	2.547178	2.227281	0.801139
min	7	7.833333	-2.66667
25%	11	10.83333	-0.66667
50%	12	12.33333	-0.16667
75%	14	14	0.5
max	19	18.16667	3



THANK YOU FOR LISTENING!