High School Student Performance & Demographics

MATH 4780

Kayley Reith, Patrick Campbell, Jake Konrad

IMPORTANCE

16.9 M

High school students (2021)

6%

High school students don’t graduate

414,000

Students each year don’t complete high school

OBJECTIVE

Comprehensively examine the relationships between student final math grade performance and independent variables, to further assess:

- The strength and direction of associations
- Identify influential points and run diagnostics
- Gain insights into nuanced patterns that contribute to student final grade performance

DATA

395 High School Students

Age 14-21

32 Independent Variables

- Study time
- Parent education
- Social time
- Extra paid classes
- Etc.

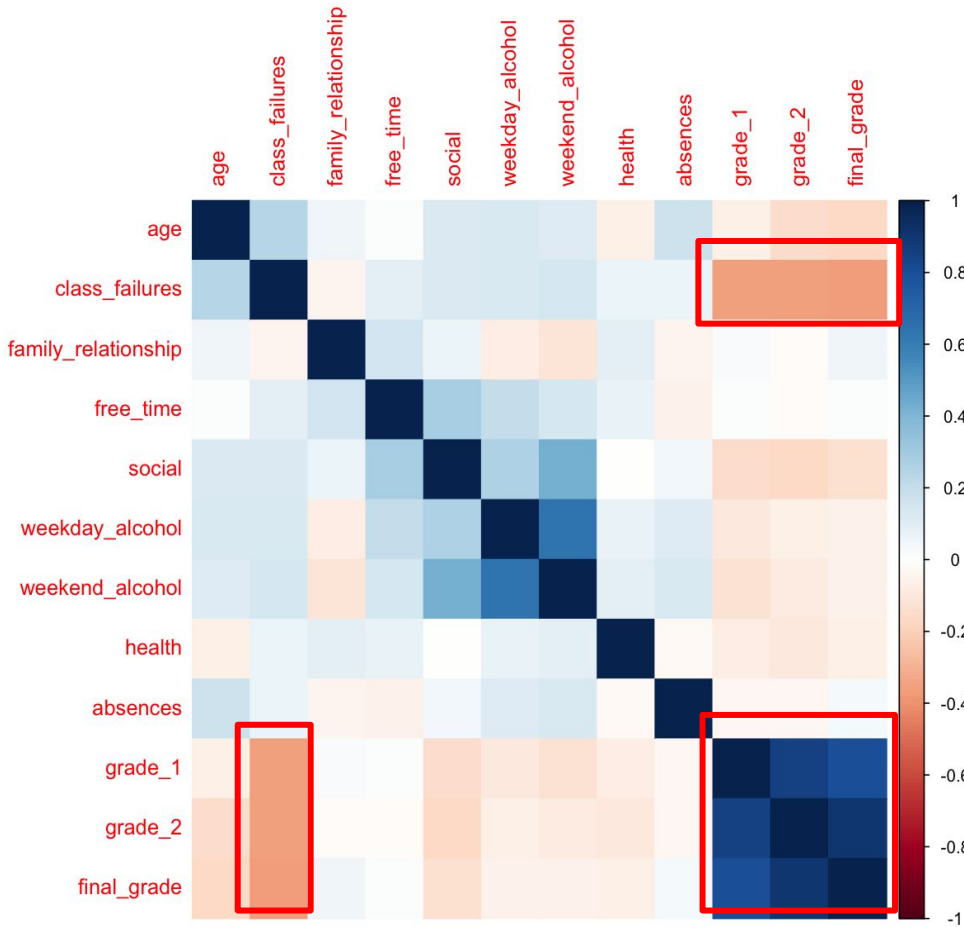
Data Structure

father_education	mother_job	father_job	school_choice_reason	guardian	travel_time	study_time	class_failures	school_support	family_support	extra_paid_classes
higher education	at_home	teacher	course	mother	15 to 30 min.	2 to 5 hours	0	yes	no	no
primary education (4th grade)	at_home	other	course	father	<15 min.	2 to 5 hours	0	no	yes	no
primary education (4th grade)	at_home	other	other	mother	<15 min.	2 to 5 hours	3	yes	no	yes
5th to 9th grade	health	services	home	mother	<15 min.	5 to 10 hours	0	no	yes	yes
secondary education	other	other	home	father	<15 min.	2 to 5 hours	0	no	yes	yes
secondary education	services	other	reputation	mother	<15 min.	2 to 5 hours	0	no	yes	yes
5th to 9th grade	other	other	home	mother	<15 min.	2 to 5 hours	0	no	no	no
higher education	other	teacher	home	mother	15 to 30 min.	2 to 5 hours	0	yes	yes	no
5th to 9th grade	services	other	home	mother	<15 min.	2 to 5 hours	0	no	yes	yes
higher education	other	other	home	mother	<15 min.	2 to 5 hours	0	no	yes	yes
higher education	teacher	health	reputation	mother	<15 min.	2 to 5 hours	0	no	yes	yes
primary education (4th grade)	services	other	reputation	father	30 min. to 1 hour	5 to 10 hours	0	no	yes	no
higher education	health	services	course	father	<15 min.	<2 hours	0	no	yes	yes
secondary education	teacher	other	course	mother	15 to 30 min.	2 to 5 hours	0	no	yes	yes
5th to 9th grade	other	other	home	other	<15 min.	5 to 10 hours	0	no	yes	no
higher education	health	other	home	mother	<15 min.	<2 hours	0	no	yes	no
higher education	services	services	reputation	mother	<15 min.	5 to 10 hours	0	no	yes	yes
secondary education	other	other	reputation	mother	30 min. to 1 hour	2 to 5 hours	0	yes	yes	no
5th to 9th grade	services	services	course	mother	<15 min.	<2 hours	3	no	yes	no
secondary education	health	other	home	father	<15 min.	<2 hours	0	no	no	yes
secondary education	teacher	other	reputation	mother	<15 min.	2 to 5 hours	0	no	no	no
higher education	health	health	other	father	<15 min.	<2 hours	0	no	yes	yes
5th to 9th grade	teacher	other	course	mother	<15 min.	2 to 5 hours	0	no	no	no
5th to 9th grade	other	other	reputation	mother	15 to 30 min.	2 to 5 hours	0	no	yes	no
higher education	services	health	course	mother	<15 min.	5 to 10 hours	0	yes	yes	yes
5th to 9th grade	services	services	home	mother	<15 min.	<2 hours	2	no	yes	yes
5th to 9th grade	other	other	home	mother	<15 min.	<2 hours	0	no	yes	yes
5th to 9th grade	health	services	other	mother	<15 min.	<2 hours	0	no	no	yes
higher education	services	other	home	mother	<15 min.	2 to 5 hours	0	yes	yes	no
higher education	teacher	teacher	home	mother	<15 min.	2 to 5 hours	0	no	yes	yes
higher education	health	services	home	mother	<15 min.	2 to 5 hours	0	no	yes	yes

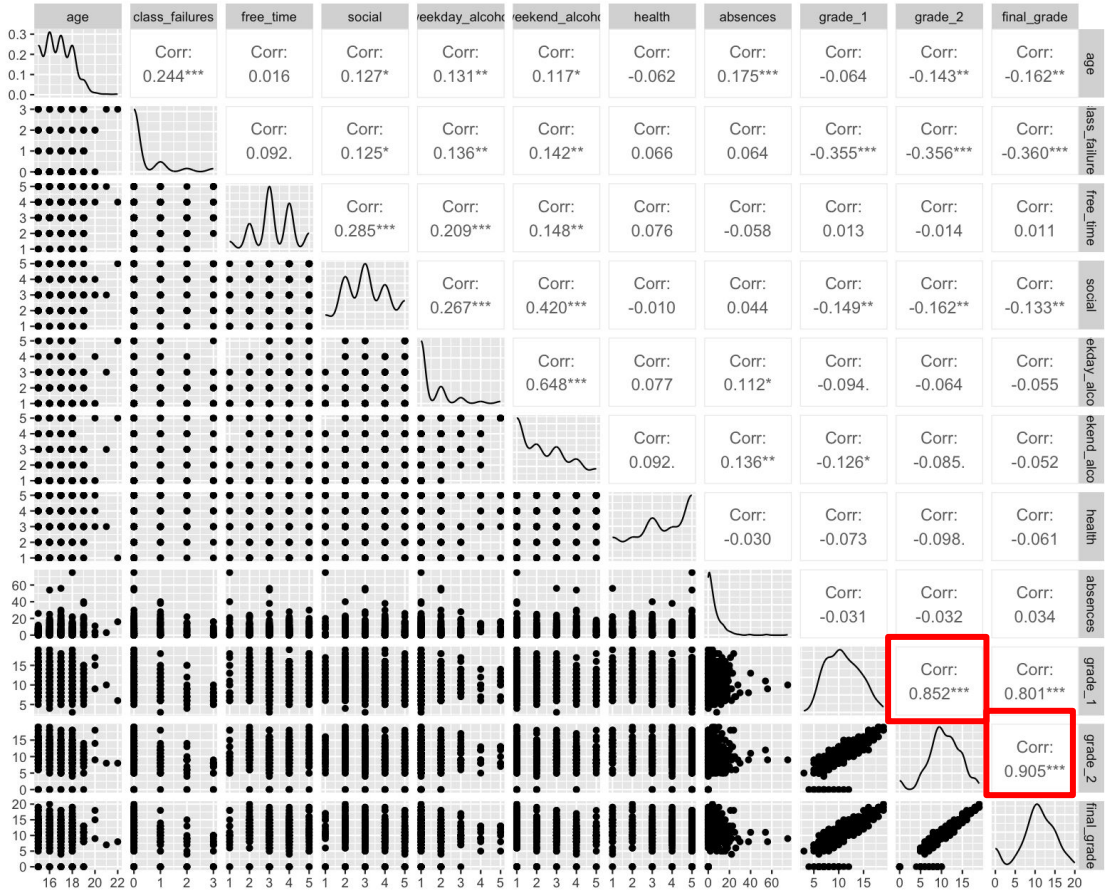
- Mixed data (Ordinal and Numeric)
- Mutate and Factorize Variables
- Variable Transformations

EXPLORATORY VISUALIZATION

The most influential variables correlated to students' success of their final math grade was the scores of **Test 1** and **Test 2**. However, they're collinear.

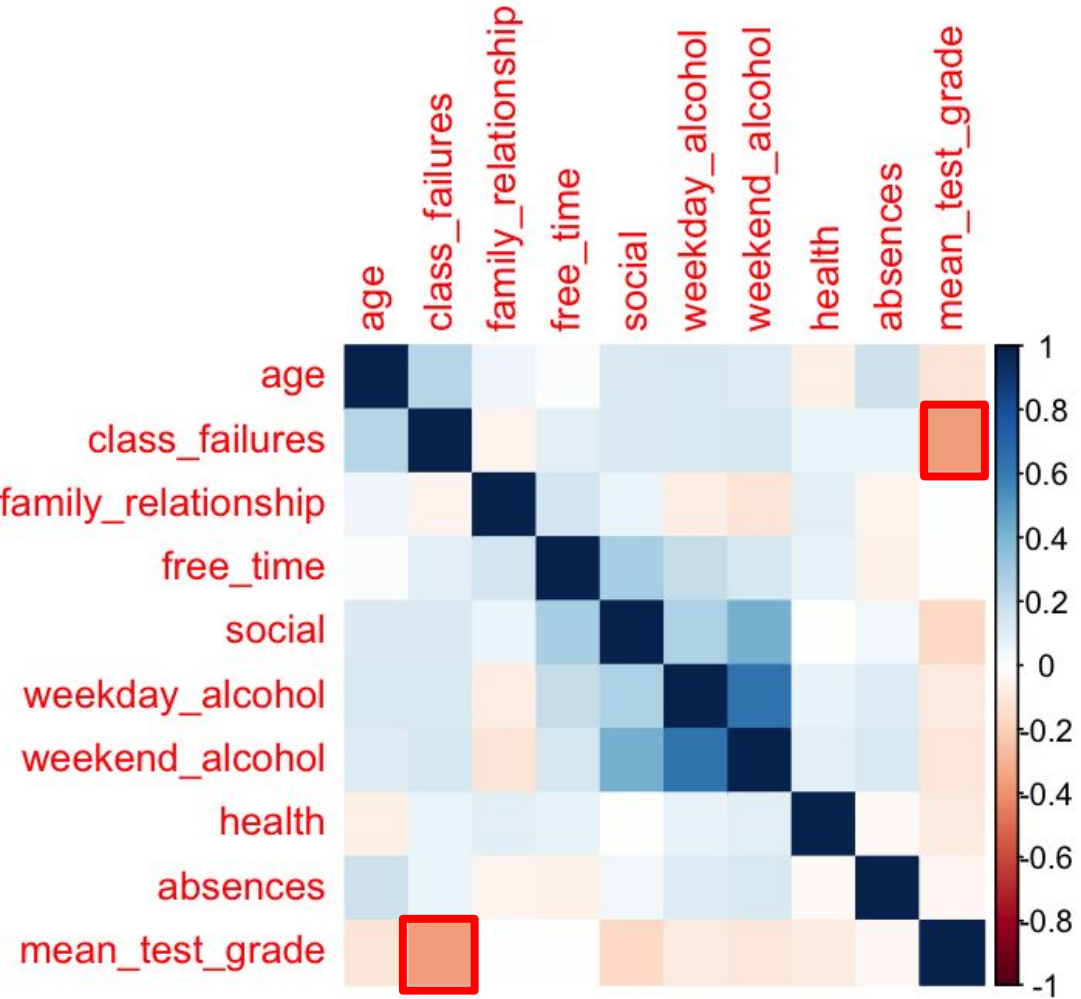


EXPLORATORY VISUALIZATION



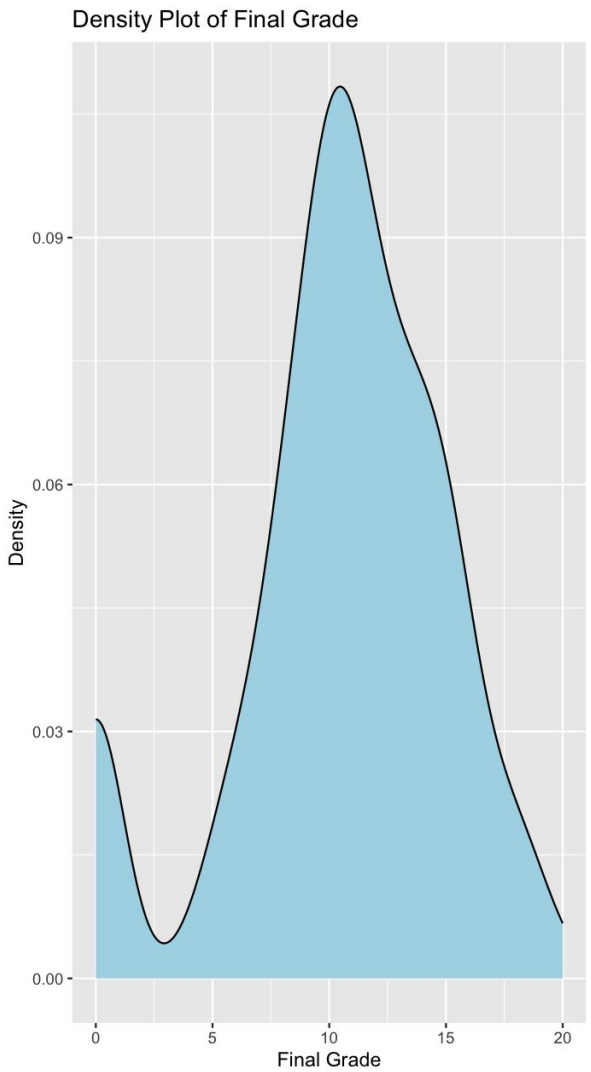
Positive Correlations:

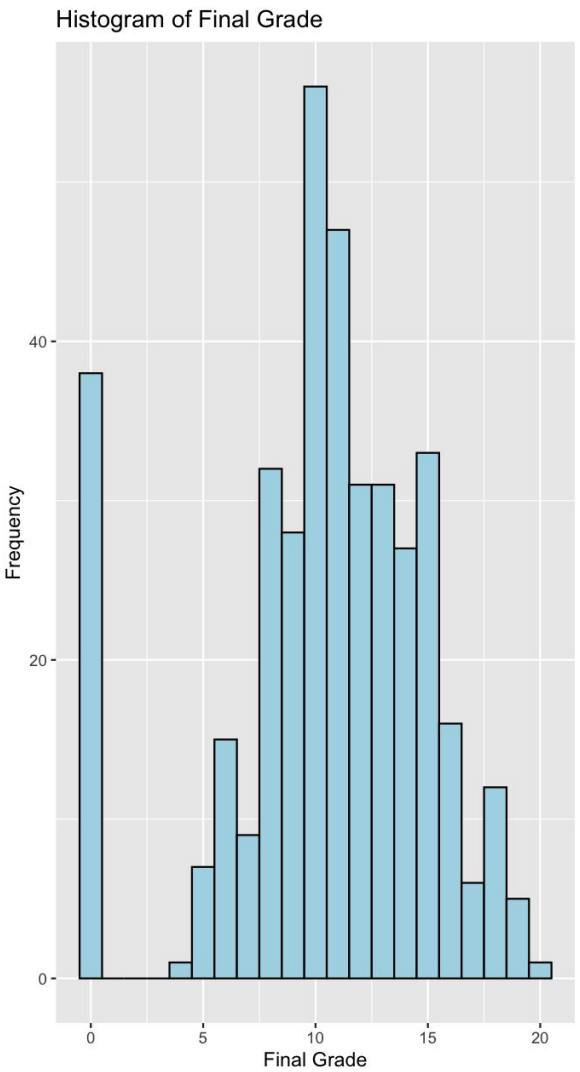
- Test Grade 1
- Test Grade 2
- Linearly independent variables are correlated
- Modification create new mean test grade



Average Test Grade between Test 1 and Test 2 :

- Yields Class Failures to be somewhat influential in the performance of final grades



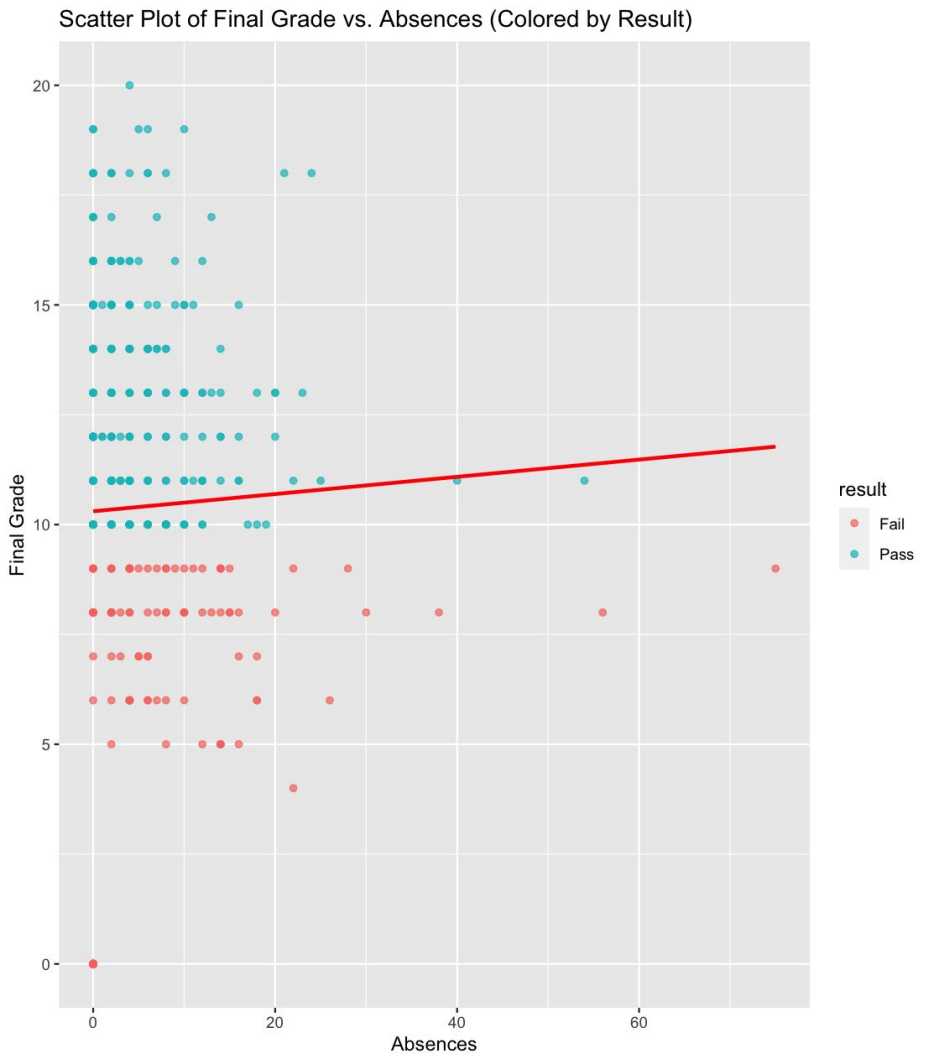


EXPLORATORY VISUALIZATION

- Mean Final Grade: 10.1
- Range: 0-20
- Bimodal distribution
- Majority of students scored between 10 and 12 points
- Second node displays 38 students scoring zero

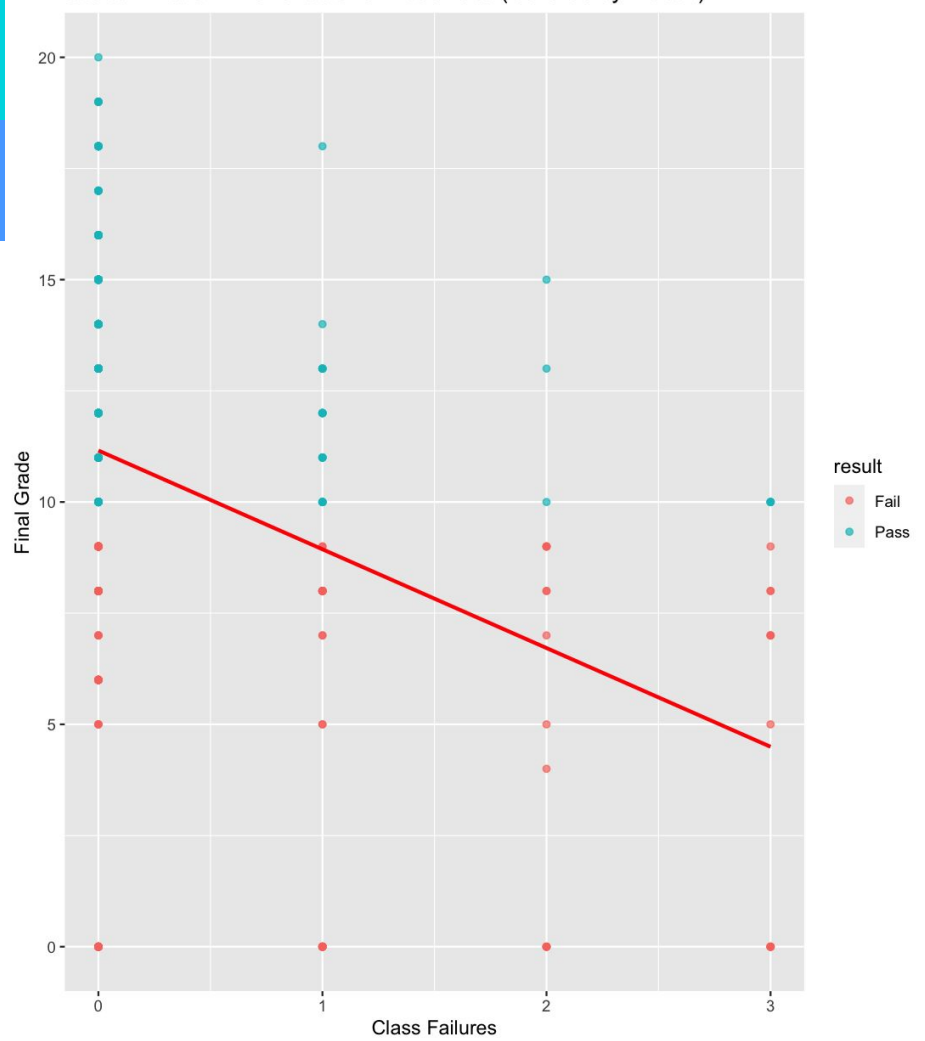
EXPLORATORY VISUALIZATION

Class Absences don’t seem to have an influence on whether students pass or fail the class



EXPLORATORY VISUALIZATION

Negative correlation between final grade and class failures. This means that students with more class failures tend to have lower final grades and **fail**.



MODEL SELECTION

Multiple Linear Regression:

- Quantify Variable Importance
- Prediction Accuracy
- Variable Selection (Stepwise Regression)

Classification Tree:

- Easy Interpretation for Stakeholders
- No Assumption of Linearity
- Inherent Variable Importance Ranking
- Generated Based on Purest Subsets
- Predicts Class (Pass/Fail)

Categorical Variables

```
educ_data <- educ_data %>%
  mutate(mother_4education = as.factor(ifelse(mother_education == 'primary education (4th grade)', 1, 0))) %>%
  mutate(mother_5to9education = as.factor(ifelse(mother_education == '5th to 9th grade', 1, 0))) %>%
  mutate(mother_secondaryeducation = as.factor(ifelse(mother_education == 'secondary education', 1, 0))) %>%
  mutate(mother_highereducation = as.factor(ifelse(mother_education == 'higher education', 1, 0))) %>%
  mutate(father_4education = as.factor(ifelse(father_education == 'primary education (4th grade)', 1, 0))) %>%
  mutate(father_5to9education = as.factor(ifelse(father_education == '5th to 9th grade', 1, 0))) %>%
  mutate(father_secondaryeducation = as.factor(ifelse(father_education == 'secondary education', 1, 0))) %>%
  mutate(father_highereducation = as.factor(ifelse(father_education == 'higher education', 1, 0))) %>%
  mutate(extraPaidClasses = as.factor(ifelse(extra_paid_classes == 'yes', 1, 0))) %>%
  mutate(Activities = as.factor(ifelse(activities == 'yes', 1, 0))) %>%
  mutate(study_time2to5 = as.factor(ifelse(study_time  == '2 to 5 hours', 1, 0))) %>%
  mutate(study_time5to10 = as.factor(ifelse(study_time == '5 to 10 hours', 1, 0))) %>%
  mutate(study_time10more = as.factor(ifelse(study_time == '>10 hours', 1, 0)))
```

- Many categorical variables created for these multi-leveled factor variables
- Not all categorical variables included in the models to prevent overfitting

Original Model

```
lm1 <- lm(final_grade ~ mean_test_grade + absences + class_failures + mother_4education + mother_5to9education + mother_secondaryeducation +
mother_highereducation + father_4education + father_5to9education + father_secondaryeducation + father_highereducation + extraPaidClasses + Activities
+ study_time10more, data = educ_data)
```

We used most of the variables available in the dataset, with a couple exceptions

Output

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.46055    1.96301  -0.235   0.8146
mean_test_grade             1.18675    0.03347  35.460   <2e-16 ***
absences                    0.03394    0.01320   2.572   0.0105 *
class_failures             -0.30304    0.15799  -1.918   0.0558 .
mother_4education1         -1.25347    1.23007  -1.019   0.3088
mother_5to9education1      -1.24305    1.21768  -1.021   0.3080
mother_secondaryeducation1 -0.81328    1.22394  -0.664   0.5068
mother_highereducation1    -0.83375    1.23416  -0.676   0.4997
father_4education1         -0.44919    1.49196  -0.301   0.7635
father_5to9education1      -1.25466    1.48817  -0.843   0.3997
father_secondaryeducation1 -0.78729    1.49003  -0.528   0.5976
father_highereducation1    -1.34087    1.49482  -0.897   0.3703
extraPaidClasses1           0.27165    0.21449   1.266   0.2061
Activities1                -0.29366    0.21143  -1.389   0.1657
study_time10more1          -0.68412    0.42270  -1.618   0.1064
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.053 on 380 degrees of freedom
Multiple R-squared:  0.8064,    Adjusted R-squared:  0.7993
F-statistic: 113.1 on 14 and 380 DF,  p-value: < 2.2e-16
```

Estimates seemed contrary to what one would assume

VARIABLE SELECTION

OLS Best Subset Selection

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.7917	0.7912	0.7897	17.7775	1708.6287	587.5302	1720.5654	1731.1435	4.4048	0.0112	0.2104
2	0.7957	0.7947	0.7926	11.9074	1702.9497	581.8994	1718.8653	1702.1563	4.3420	0.0110	0.2074
3	0.7978	0.7963	0.7934	9.8410	1700.9228	579.9242	1720.8172	1689.2115	4.3197	0.0110	0.2063
4	0.7994	0.7974	0.7941	8.7005	1699.7844	578.8560	1723.6578	1680.1518	4.3073	0.0109	0.2057
5	0.8011	0.7985	0.7939	7.3820	1698.4408	577.6210	1726.2930	1670.2832	4.2927	0.0109	0.2050
6	0.8025	0.7995	0.7943	6.6212	1697.6374	576.9418	1729.4685	1662.7565	4.2840	0.0109	0.2046
7	0.8035	0.8000	0.7933	6.5937	1697.5659	576.9936	1733.3759	1658.3444	4.2832	0.0109	0.2046
8	0.8045	0.8004	0.7935	6.8065	1697.7309	577.2927	1737.5197	1654.9455	4.2850	0.0109	0.2047
9	0.8051	0.8006	0.793	7.4744	1698.3575	578.0489	1742.1252	1653.4962	4.2918	0.0109	0.2050
10	0.8057	0.8006	0.7925	8.3575	1699.2024	579.0254	1746.9490	1652.9725	4.3010	0.0109	0.2054
11	0.8061	0.8005	0.7921	9.5450	1700.3599	580.3072	1752.0854	1653.7687	4.3137	0.0110	0.2060
12	0.8062	0.8001	0.7912	11.4578	1702.2694	582.2988	1757.9738	1657.7293	4.3346	0.0110	0.2070
13	0.8063	0.7997	0.7906	13.0906	1703.8880	584.0218	1763.5713	1660.4878	4.3524	0.0111	0.2079
14	0.8064	0.7993	0.7897	15.0000	1705.7938	586.0135	1769.4560	1664.4720	4.3735	0.0111	0.2089

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp
APC: Amemiya Prediction Criteria

Model 7:
Adjusted R-Square: 0.8000
Mallow's C(p): 6.5937

Model 9:
Adjusted R-Square: 0.8006
Mallow's C(p): 7.4744

Model Predictors

Response Variable: Final Grade

Model 7 Predictors:

- mean_test_grade
- absences
- class_failures
- father_5to9education
- father_highereducation
- Activities
- study_time10more

Model 9 Predictors:

- mean_test_grade
- absences
- Class_failures
- mother_5to9education
- father_5to9education
- Father_highereducation
- extraPaidClasses
- Activities
- study_time10more

Problems

Influential Points

Affect the interpretation of the coefficients and statistical significance

Data Quality

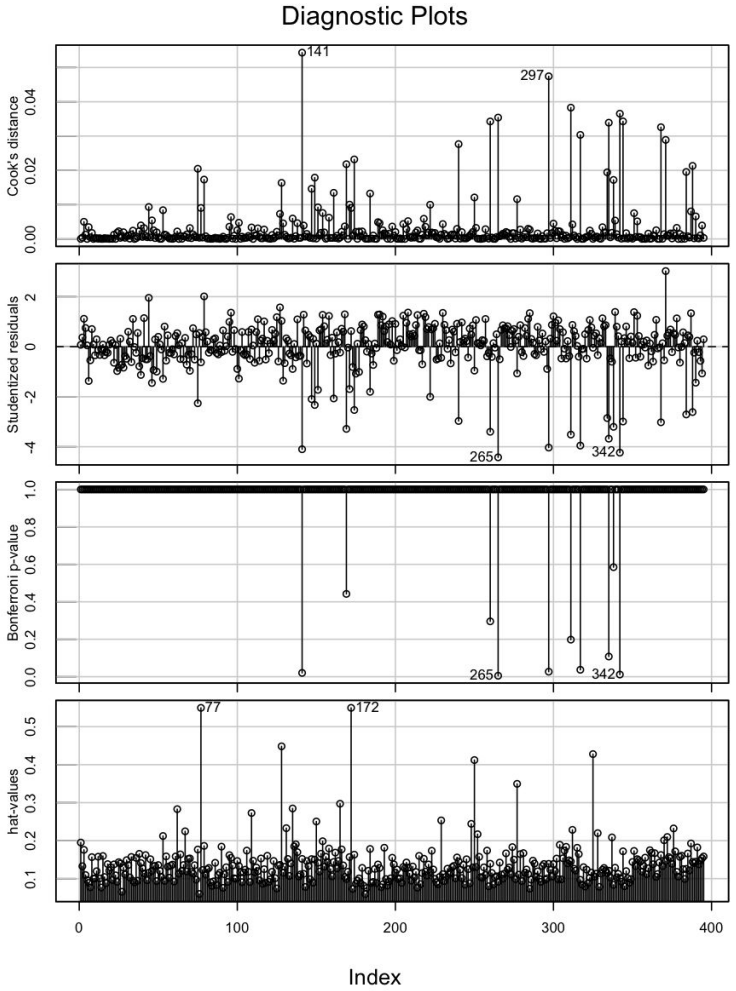
Several reported values of 0 as the final score students received.

Non-normality

Skewed and heavy-tailed errors give rise to outliers, corrupting our model.

DIAGNOSTICS

- We wanted to see if individual data points were influencing this strange output
- Influence measures helped us decide what to remove before retesting the model
- Normality, variance, and residual checks help us assess the necessity of variable transformations



Removing Outliers for Model 7

```
> summary(inf_educ1)
Potentially influential observations
```

	dfb.1_	dfb.mn__	dfb.absn	dfb.cls_	dfb.f_59	dfb.ft_1	dfb.Act1	dfb.s_10	dffit	cov.r	cook.d	hat
48	-0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.02	1.08_*	0.00	0.05
67	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	-0.01	-0.02	1.07_*	0.00	0.05
68	0.01	-0.01	0.00	0.00	0.00	0.00	-0.01	0.02	0.02	1.07_*	0.00	0.05
70	0.02	-0.03	-0.02	-0.01	0.03	0.01	0.03	-0.08	-0.10	1.07_*	0.00	0.05
71	0.00	0.02	-0.01	0.00	-0.03	-0.01	-0.03	0.09	0.10	1.07_*	0.00	0.05
75	0.05	-0.01	-0.38	0.05	0.05	0.06	-0.06	0.01	-0.40	1.11_*	0.02	0.10_*
77	0.00	0.00	-0.01	0.00	0.01	0.01	-0.01	-0.04	-0.04	1.06_*	0.00	0.04
78	0.01	-0.01	-0.02	-0.01	0.04	0.01	-0.03	0.11	0.13	1.06_*	0.00	0.05
106	0.02	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	0.07	0.08	1.07_*	0.00	0.05
122	-0.02	0.01	0.00	0.00	0.02	0.00	0.01	0.05	0.06	1.07_*	0.00	0.05
141	-0.31	0.21	0.10	0.10	0.19	0.06	0.22	-0.67	-0.78_*	0.84_*	0.07	0.05
145	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.07_*	0.00	0.04
147	0.02	0.01	0.10	-0.36	-0.17	-0.06	0.07	-0.01	-0.45_*	0.98	0.03	0.05
149	-0.25	0.22	0.09	0.10	-0.01	-0.24	0.15	0.00	-0.39	0.91_*	0.02	0.02
154	0.02	-0.02	-0.02	0.05	0.03	0.01	-0.01	0.00	0.07	1.07_*	0.00	0.05
161	-0.05	0.05	0.10	-0.23	0.08	0.06	-0.13	0.01	-0.36	0.93_*	0.02	0.02
169	-0.23	0.20	0.09	0.13	-0.19	-0.02	0.11	0.04	-0.35	0.92_*	0.02	0.02
174	0.07	-0.06	0.14	-0.51	0.10	0.06	-0.17	0.00	-0.60_*	0.90_*	0.04	0.04
184	0.02	0.04	-0.42	0.07	0.05	0.06	-0.07	0.01	-0.44_*	1.11_*	0.02	0.11_*
211	0.03	-0.03	0.02	-0.01	-0.02	-0.01	0.02	0.10	0.12	1.06_*	0.00	0.05
240	-0.14	0.12	0.11	-0.07	-0.20	-0.03	0.11	0.03	-0.34	0.90_*	0.01	0.02
257	-0.02	0.00	0.01	0.00	0.03	0.01	0.02	0.08	0.10	1.06_*	0.00	0.04
260	-0.07	0.19	0.11	0.05	-0.30	-0.08	-0.19	-0.72	-0.87_*	0.78_*	0.09	0.05
265	-0.10	0.14	0.15	0.13	-0.33	0.02	-0.24	0.10	-0.52_*	0.73_*	0.03	0.02
272	0.01	-0.01	0.00	0.00	0.01	0.01	-0.01	-0.04	-0.05	1.07_*	0.00	0.04
277	0.05	0.04	-0.63	0.09	-0.10	-0.01	0.06	0.00	-0.65_*	1.24_*	0.05	0.20_*
283	0.00	0.00	-0.01	0.00	-0.02	-0.01	0.01	0.06	0.07	1.06_*	0.00	0.04
294	0.03	-0.06	-0.01	-0.01	0.03	0.02	0.04	-0.10	-0.13	1.07_*	0.00	0.05
297	-0.13	0.17	0.15	0.08	-0.03	-0.35	-0.18	0.03	-0.52_*	0.72_*	0.03	0.02
299	0.00	0.01	-0.01	0.00	-0.02	-0.01	-0.02	0.05	0.06	1.07_*	0.00	0.05
304	-0.06	0.05	-0.02	0.02	0.04	0.00	0.03	0.12	0.15	1.06_*	0.00	0.05
308	0.00	0.01	-0.06	-0.01	0.00	-0.03	0.02	-0.01	-0.08	1.08_*	0.00	0.06
311	0.00	0.06	0.15	-0.15	-0.30	-0.02	-0.23	0.07	-0.48_*	0.79_*	0.03	0.02
317	-0.24	0.17	0.13	0.15	0.14	0.16	-0.18	0.07	-0.40	0.80_*	0.02	0.01
334	-0.15	0.19	0.12	0.13	-0.26	0.01	-0.19	0.08	-0.43_*	0.84_*	0.02	0.02

We modified the dataset for Model 7 by removing influential points to see if that was affecting results

42 influential points in this model, out of 395 total

Removing Outliers for Model 9

```
> summary(inf_educ1)
Potentially influential observations
```

	dfb.1_	dfb.mn__	dfb.absn	dfb.cls_	dfb.m_59	dfb.f_59	dfb.ft_1	dfb.ePC1	dfb.Act1	dfb.s_10	dffit	cov.r	cook.d	hat
48	-0.01	0.02	0.00	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.02	0.03	1.09_*	0.00	0.06
67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.09_*	0.00	0.05
68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.08_*	0.00	0.05
70	0.01	-0.03	-0.02	0.00	0.01	0.02	0.01	0.02	0.03	-0.07	-0.09	1.08_*	0.00	0.06
75	0.07	-0.01	-0.41	0.04	0.01	0.06	0.07	-0.07	-0.07	0.01	-0.44	1.10_*	0.02	0.11_*
122	-0.02	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.01	0.05	0.07	1.08_*	0.00	0.05
141	-0.36	0.21	0.10	0.14	0.13	0.15	0.07	0.22	0.22	-0.68	-0.82_*	0.80_*	0.07	0.05
145	0.01	-0.01	-0.01	0.02	0.01	-0.01	0.00	0.00	-0.01	0.00	0.03	1.08_*	0.00	0.05
149	-0.28	0.22	0.09	0.12	0.04	-0.02	-0.23	0.14	0.15	-0.01	-0.41	0.89_*	0.02	0.03
154	0.02	-0.02	-0.02	0.04	-0.02	0.03	0.01	-0.01	-0.01	0.00	0.07	1.08_*	0.00	0.05
169	-0.16	0.19	0.09	0.10	-0.16	-0.15	-0.04	-0.13	0.11	0.05	-0.40	0.90_*	0.02	0.03
174	0.05	-0.05	0.14	-0.50	0.08	0.08	0.07	0.02	-0.17	0.00	-0.61_*	0.87_*	0.04	0.04
184	-0.01	0.04	-0.40	0.08	0.02	0.04	0.05	0.07	-0.06	0.00	-0.42	1.12_*	0.02	0.11_*
240	-0.12	0.10	0.09	-0.05	-0.17	-0.16	-0.06	0.10	0.11	0.02	-0.38	0.89_*	0.01	0.02
260	0.02	0.17	0.09	0.01	-0.24	-0.24	-0.11	-0.17	-0.19	-0.71	-0.92_*	0.73_*	0.08	0.05
265	0.00	0.12	0.14	0.08	-0.26	-0.27	-0.01	-0.22	-0.24	0.12	-0.62_*	0.67_*	0.04	0.02
272	0.01	-0.01	0.00	0.00	-0.02	0.01	0.00	-0.01	-0.01	-0.03	-0.04	1.08_*	0.00	0.05
277	0.02	0.04	-0.60	0.10	0.04	-0.11	0.00	0.09	0.06	-0.01	-0.63_*	1.25_*	0.04	0.21_*
297	-0.09	0.18	0.16	0.05	0.07	-0.03	-0.33	-0.23	-0.20	0.04	-0.59_*	0.65_*	0.03	0.02
308	0.01	0.01	-0.08	-0.02	0.00	0.00	-0.04	-0.02	0.02	-0.01	-0.10	1.09_*	0.00	0.06
311	-0.07	0.07	0.16	-0.11	0.22	-0.35	0.01	0.15	-0.23	0.06	-0.55_*	0.73_*	0.03	0.02
317	-0.12	0.15	0.11	0.10	-0.30	0.21	0.12	-0.21	-0.17	0.09	-0.55_*	0.76_*	0.03	0.02
334	-0.14	0.17	0.10	0.14	-0.18	-0.22	-0.03	0.16	-0.17	0.06	-0.48	0.82_*	0.02	0.02
335	-0.08	0.16	0.08	0.08	-0.22	-0.24	-0.12	0.24	-0.17	-0.70	-0.90_*	0.75_*	0.08	0.05
338	-0.24	0.22	0.14	0.13	0.19	-0.27	0.02	-0.16	0.13	0.07	-0.50_*	0.80_*	0.02	0.02
342	-0.01	0.04	0.17	-0.22	0.05	-0.05	-0.38	0.16	-0.20	-0.01	-0.61_*	0.65_*	0.04	0.02
344	-0.09	0.06	0.12	-0.10	-0.23	-0.21	-0.07	0.14	0.15	0.04	-0.50_*	0.77_*	0.02	0.02
368	-0.19	0.13	0.12	-0.08	0.09	0.11	0.10	-0.19	0.13	0.03	-0.38	0.85_*	0.01	0.02

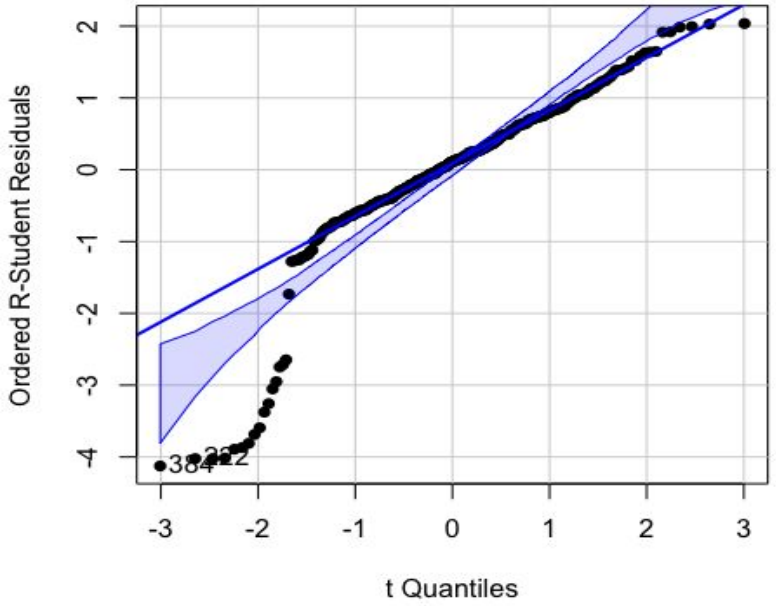
Now we modify model 9 to see how that affects estimates

28 influential points in the model out of 395 total

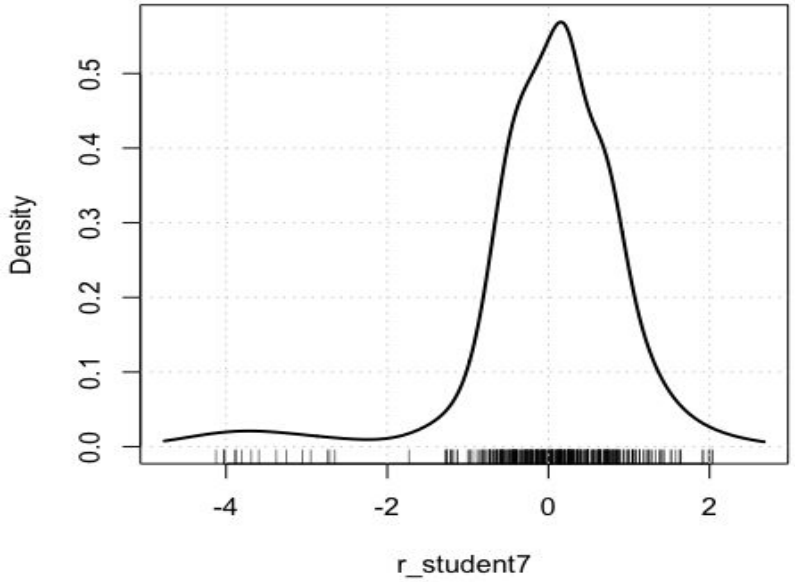
Normality Plots for Model 7

Non-normality exists here

QQPlot



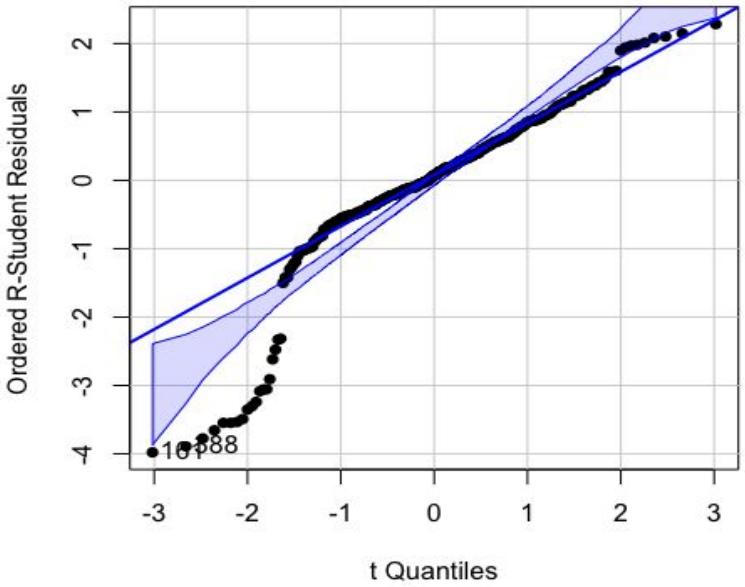
Density Plot



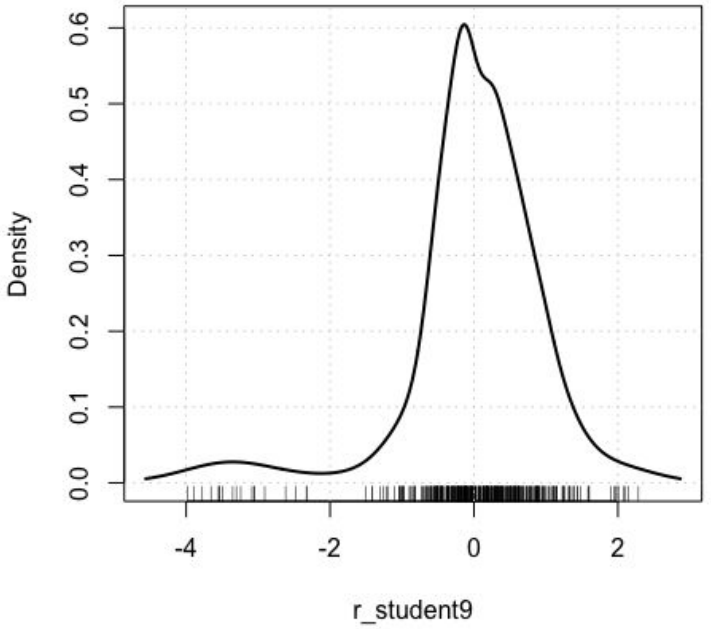
Normality Plots for Model 9

Non-normality exists here

QQPlot



Density Plot



Resolved Model 7

```
> (lambda <- pt$result[1,1])
[1] 7.072431
> bestlm7 <- lm(log(final_grade + 10)^lambda ~ mean_test_grade + absences + class_failures + father_4education + father_se
condaryeducation + Activities + study_time10more, data = better_educ_data7)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -618.980     74.067  -8.357  1.6e-15 ***
mean_test_grade            305.816      5.564  54.960  < 2e-16 ***
absences                    -2.711      2.819  -0.962   0.3369
class_failures              -9.318     26.657  -0.350   0.7269
father_4education1          35.170     46.975   0.749   0.4545
father_secondaryeducation1  49.995     42.102   1.187   0.2359
Activities1                -21.946     34.641  -0.634   0.5268
study_time10more1          208.077    124.806   1.667   0.0964 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 323.5 on 345 degrees of freedom
Multiple R-squared:  0.9106,    Adjusted R-squared:  0.9087
F-statistic: 501.8 on 7 and 345 DF,  p-value: < 2.2e-16
```

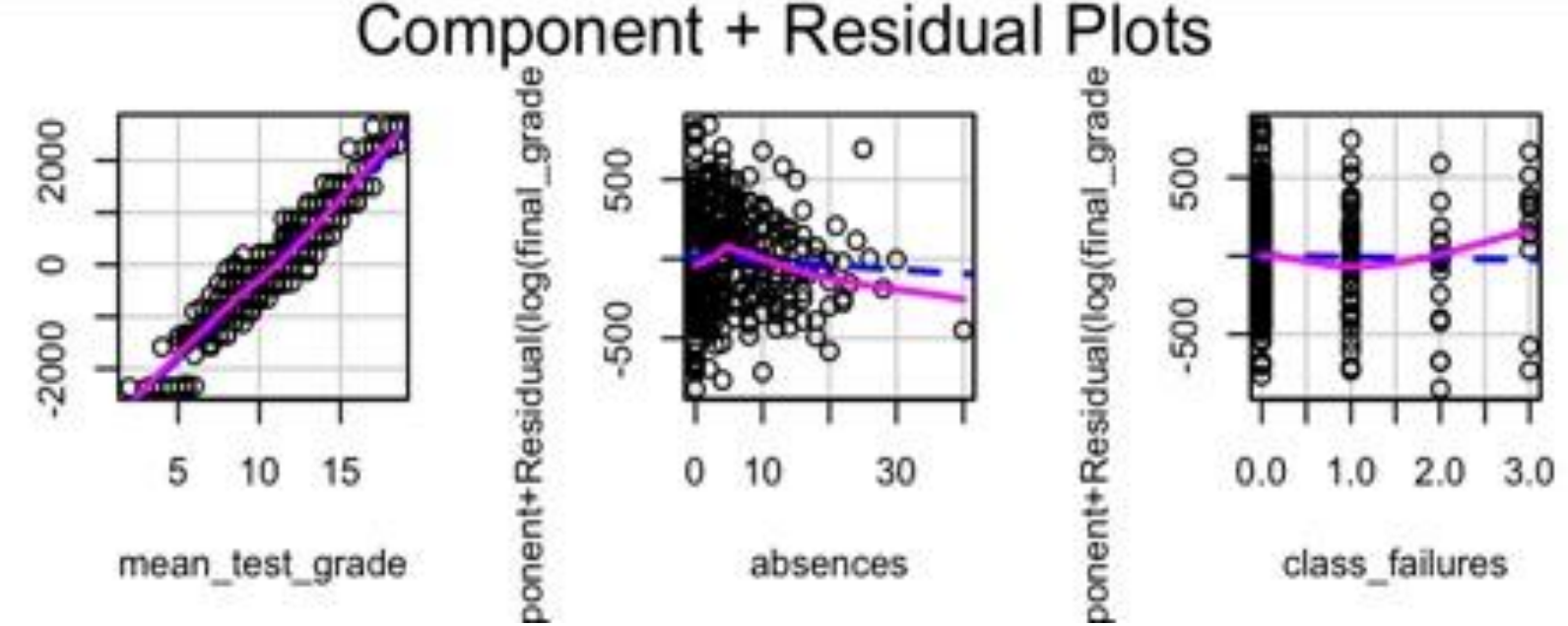
- Transform response variable using log(y + 10)^bcPower
- **91.06% of the variation** in the response variable is explained by the model
- Coefficients make more sense
- Difficult to interpret practically

Resolved Model 9

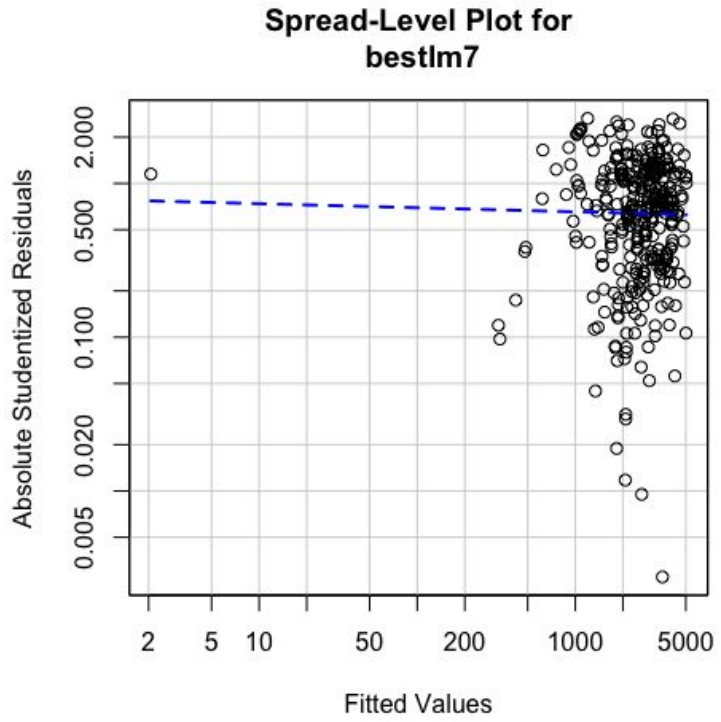
```
> (lambda <- pt$result[1,1])
[1] 7.302378
> bestlm9 <- lm(log(final_grade + 10)^lambda ~ mean_test_grade + absences + class_failures + mother_5to9education + father
_5to9education + father_highereducation + extraPaidClasses + Activities + study_time10more, data = better_educ_data9)
> summary(bestlm9)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -881.044     97.641  -9.023   <2e-16 ***
mean_test_grade        409.991      7.404  55.375   <2e-16 ***
absences                -2.201      3.786  -0.582    0.561
class_failures         -43.459     34.985  -1.242    0.215
mother_5to9education1  -41.417     55.537  -0.746    0.456
father_5to9education1  -58.137     56.465  -1.030    0.304
father_highereducation1 -59.497    59.605  -0.998    0.319
extraPaidClasses1       -3.688     47.164  -0.078    0.938
Activities1            -39.850     46.416  -0.859    0.391
study_time10more1      140.433    108.252   1.297    0.195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 440.5 on 357 degrees of freedom
Multiple R-squared:  0.9106,    Adjusted R-squared:  0.9084
F-statistic: 404.2 on 9 and 357 DF,  p-value: < 2.2e-16
```

- Transform response variable using log(y + 10)^bcPower
- **91.06% of the variation** in the response variable is explained by the model
- Coefficients make more sense, but some also don't here
- Difficult to interpret practically

Model 7 Residual Plots (Data Omitted)

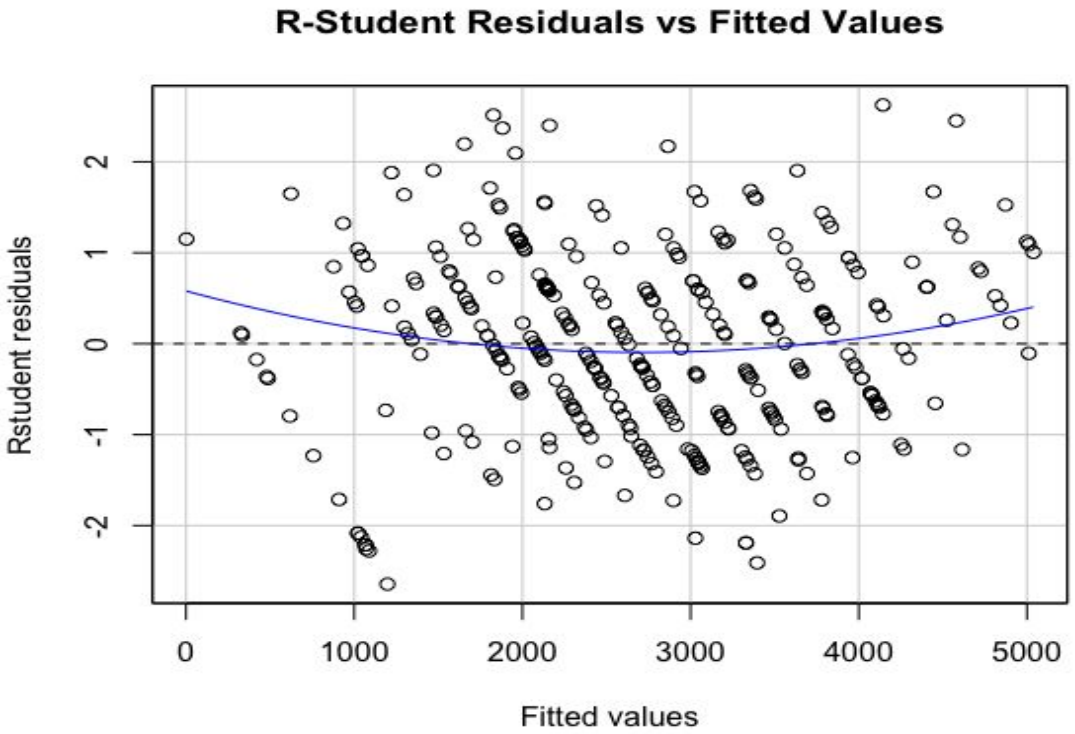


Model 7 Tukey's Spread Level Plot (Data Omitted)



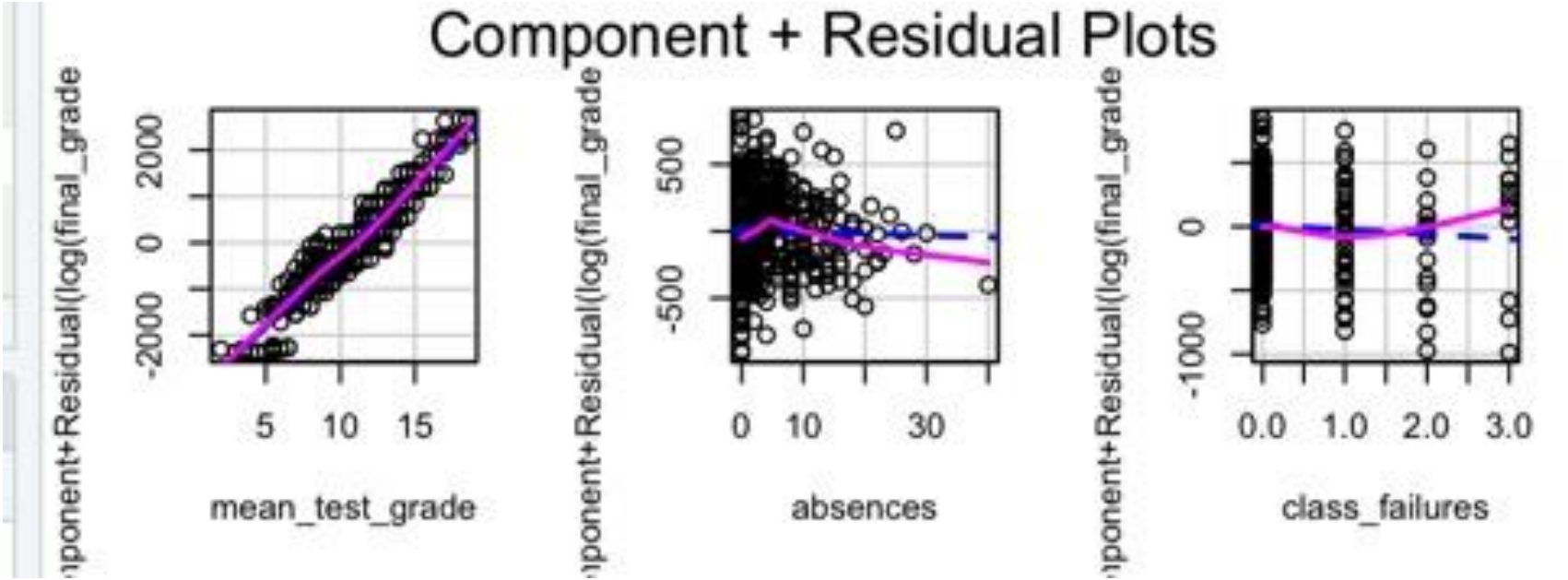

```
> car::spreadLevelPlot(bestlm7, smooth = FALSE)

Suggested power transformation:  1.02643
```

Model 7 Residual Plots (Data Omitted)

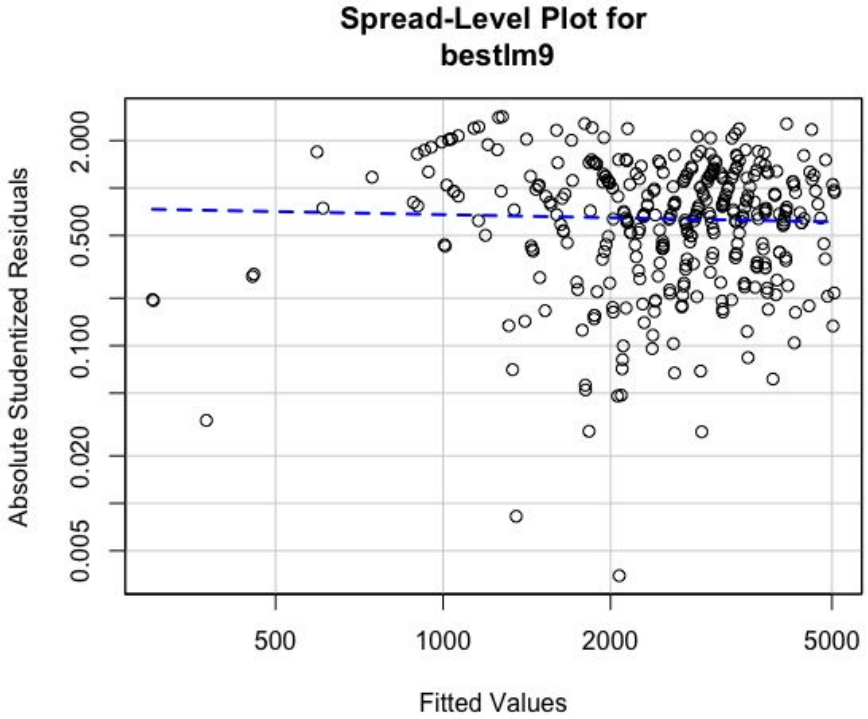


This shows no further need to transform the data

Model 9 Component + Residual Plots (Data Omitted)

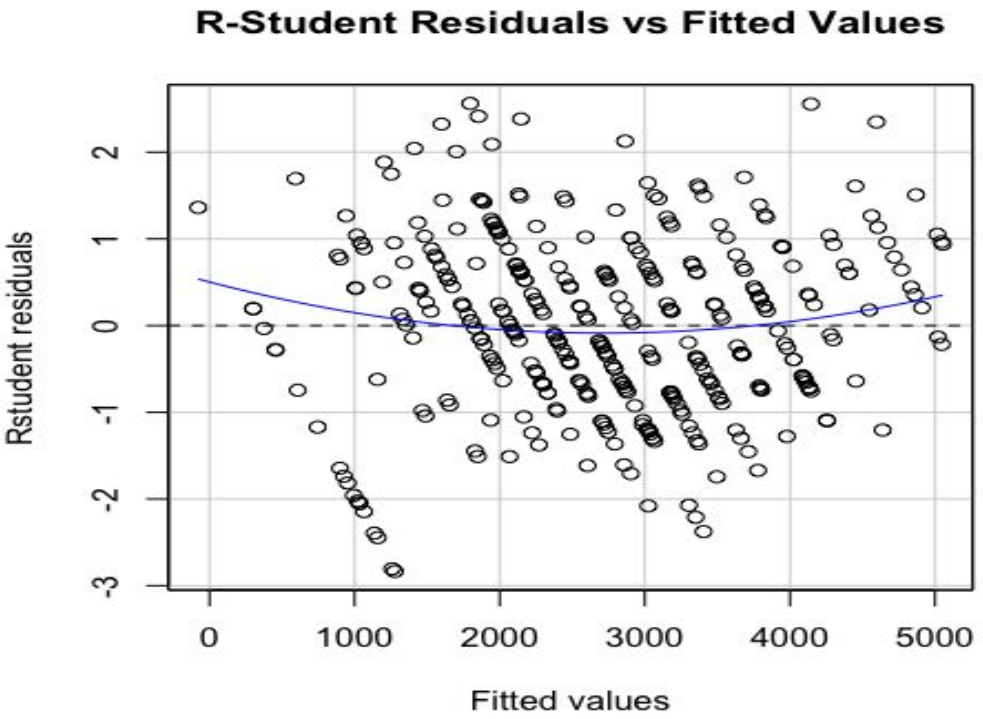


Model 9 Tukey's Spread Level Plot




```
> car::spreadLevelPlot(bestlm9, smooth = FALSE)

Suggested power transformation:  1.065
```

Model 9 Residual Plots (Data Omitted)

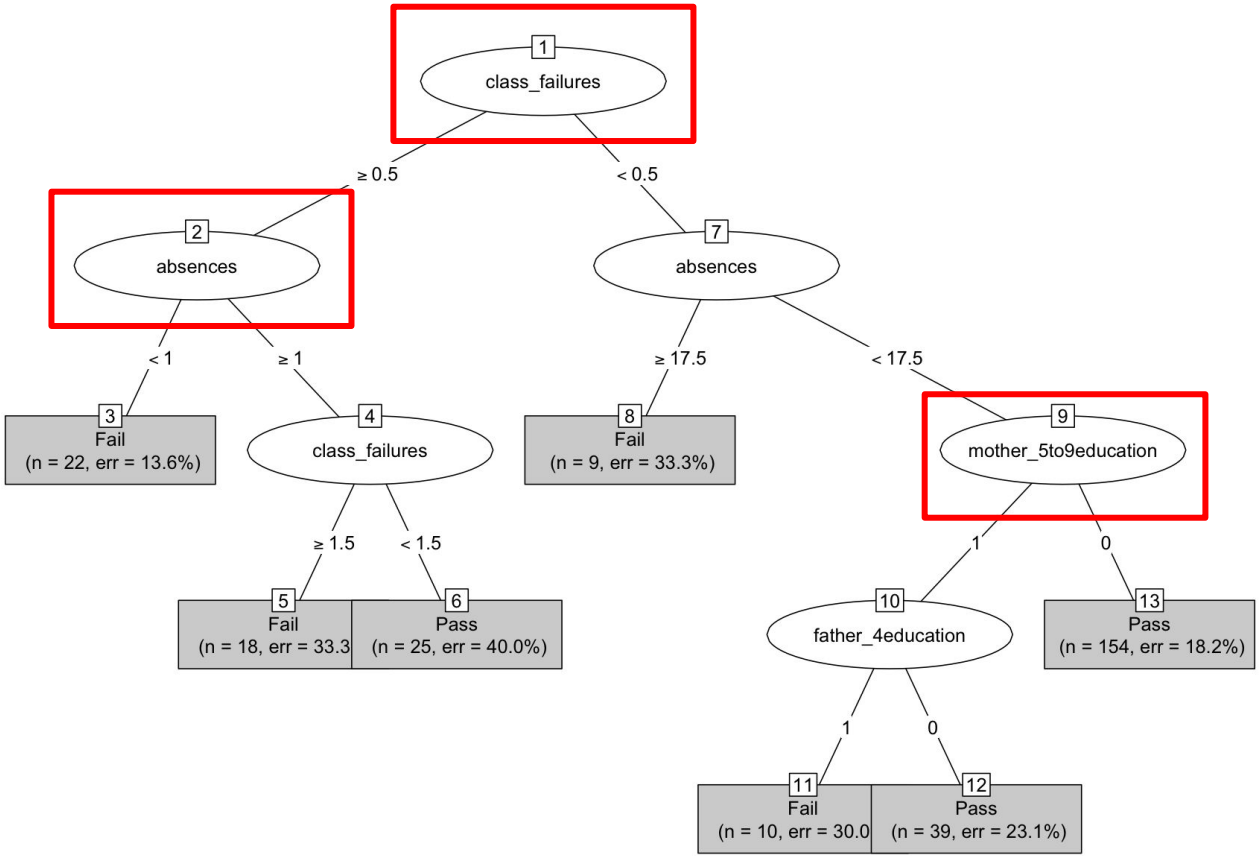


This also shows no further transformation is needed on the data

Classification Tree

```r
#testing score improvement pass fail
educ_data$result <- ifelse(educ_data$final_grade < 10, "Fail", "Pass")
set.seed(1)
myIndex <- createDataPartition(educ_data$result, p=0.7, list=FALSE) #70% training, 30% validation
trainSet <- educ_data[myIndex,]
validationSet <- educ_data[-myIndex,]
set.seed(1)
default_tree9 <- rpart(result ~  absences + class_failures + mother_4education + mother_5to9education
                       + mother_secondaryeducation + mother_highereducation + father_4education + father_5to9education
                       + father_secondaryeducation + father_highereducation + extraPaidClasses + Activities + study_time10more,
                       data = trainSet,
                       method = "class")
summary(default_tree9)
prp(default_tree9,
    type = 1,
    extra = 1,
    under = TRUE)
library(rpart.plot)
install.packages("partykit")
library(partykit) # Source: R/decision_tree_partykit.R https://parsnip.tidymodels.org/reference/details_decision_tree_partykit.html
party_obj <- as.party(default_tree9)
plot(party_obj, type = "simple")
```

CLASSIFICATION TREE



- Students with **more class failures** are more likely to **fail** the course
- Students with **more absences** are more likely to **fail** the course
- Students with **parents who have higher levels of education** are more likely to **pass** the course

RESULTS

- Model 7 proved to be the best model for measuring final grade in this math class based on adjusted R-Squared

- For this model, the most **statistically significant** predictor is mean test grade being high (positive)

- Some other factors: Studying (10 hours), higher education of parents

- **Negative impacts** on grade would be activities and parents having only up to high school education

RESULTS

- Previous class failures and absences worsen student performance **(fail)**

- Students whose parents, specifically fathers, have higher education are in better positions to **pass**

- **Activities,** although beneficial for youth development may impact grades

INDUSTRY FINDINGS

- Inspire teachers to create attendance incentives to improve student performance
- Develop more programs for high school students who are first generation students
- Provide more resources to stakeholders

LIMITATIONS

- Overfitting data
- Omitted Variable Bias
- Interpretation Issues
- Statistical Significance vs. Practical Significance

SUGGESTIONS & FUTURE DIRECTIONS

- Increase Dataset Size
- Additional Variables
- Diverse Demographics
- Incorporate Behavioral Patterns
- Interdisciplinary Research

CITATIONS

Craft, S. (2022, February 11). High School Statistics. Think Impact. https://www.thinkimpact.com/high-school-statistics/#:~:text=The%20gender%20parity%20for%20US,15%25%20college%20drop%20out%20rate

High School of America. (2022, October 7). High School Statistics in the United States. High School of America.https://www.highschoolofamerica.com/united-states-high-school-statistics/#:~:text=In%202021%2C%20approximately%2016.9%20million,enrolled%20in%20public%20high%20schools

Data: High School Student Performance & Demographics. Kaggle. 2023. https://www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics

QUESTIONS