Analytical Insights on Student Alcohol Consumption

By: Andy Ma, Haoliang (Jack) Hu, Truc Ly, Vi Tran, Shawn Ha

Table of **contents**

01

Dataset Introduction

What is/are the importance/questions?

02

Visualizations

Exploratory data analysis and descriptive statistics

03

Regression/Classification Methods

Applying regression and classification to our dataset

04

Conclusions

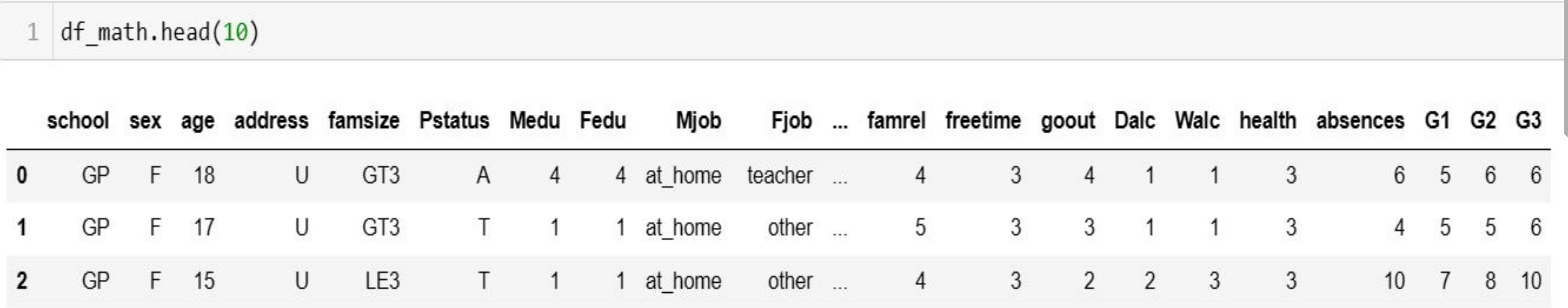
Recommendations/Impact

01

Introduction to the Dataset

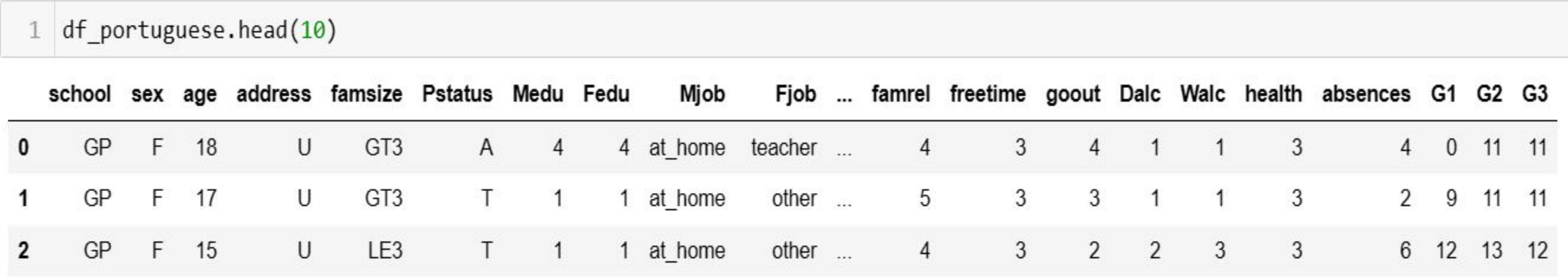
About Our Dataset

- Real-life data based on self-reports and performance metrics, from Portugal
- Our dataset consists of a total of 1,044 records of grades achieved by students in enrolled in secondary schools
- It includes classes in Math and Portuguese
- Students have a variety of attributes: age, daily alcohol consumption, etc.



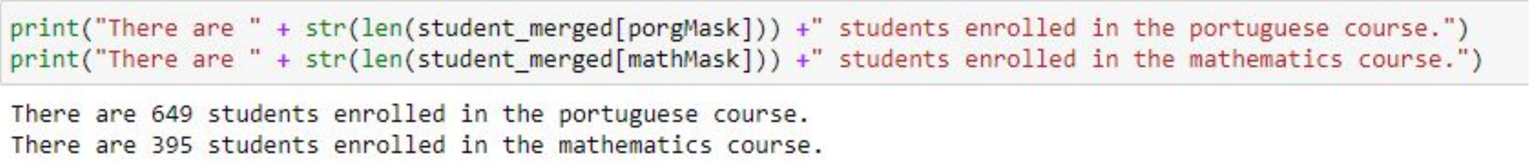
```
df_math.head(10)
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10



```
df_portuguese.head(10)
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12



```
print("There are " + str(len(student_merged[porgMask])) +" students enrolled in the portuguese course.")
print("There are " + str(len(student_merged[mathMask])) +" students enrolled in the mathematics course.")
```

```
There are 649 students enrolled in the portuguese course.
There are 395 students enrolled in the mathematics course.
```

Central Questions

- Is there a relationship between regular alcohol consumption and student class performance?
- Do other seemingly important variables affect student grades (like father’s education, extra educational support?)
- Identify any interrelationships between the student’s attributes

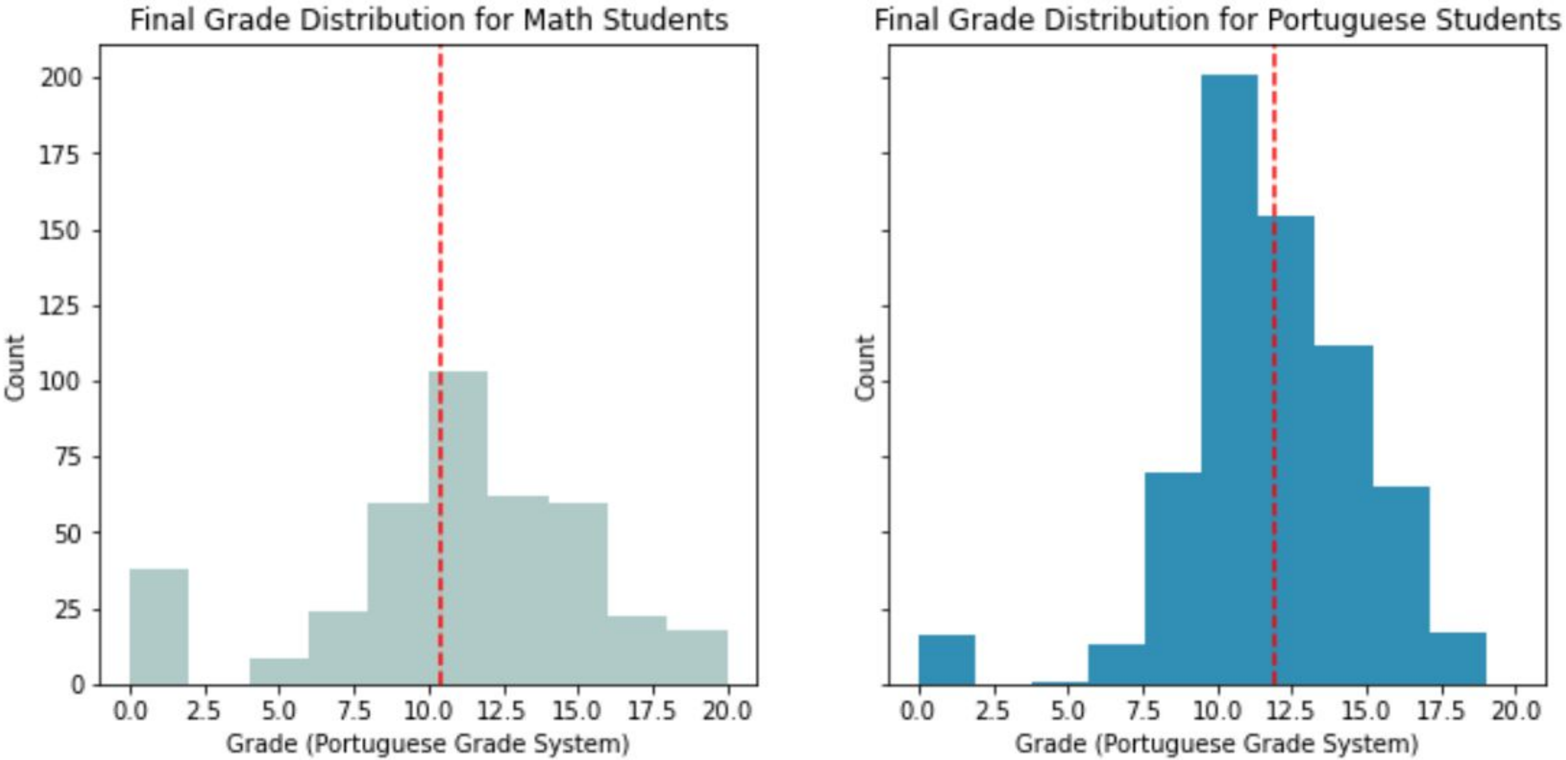
02

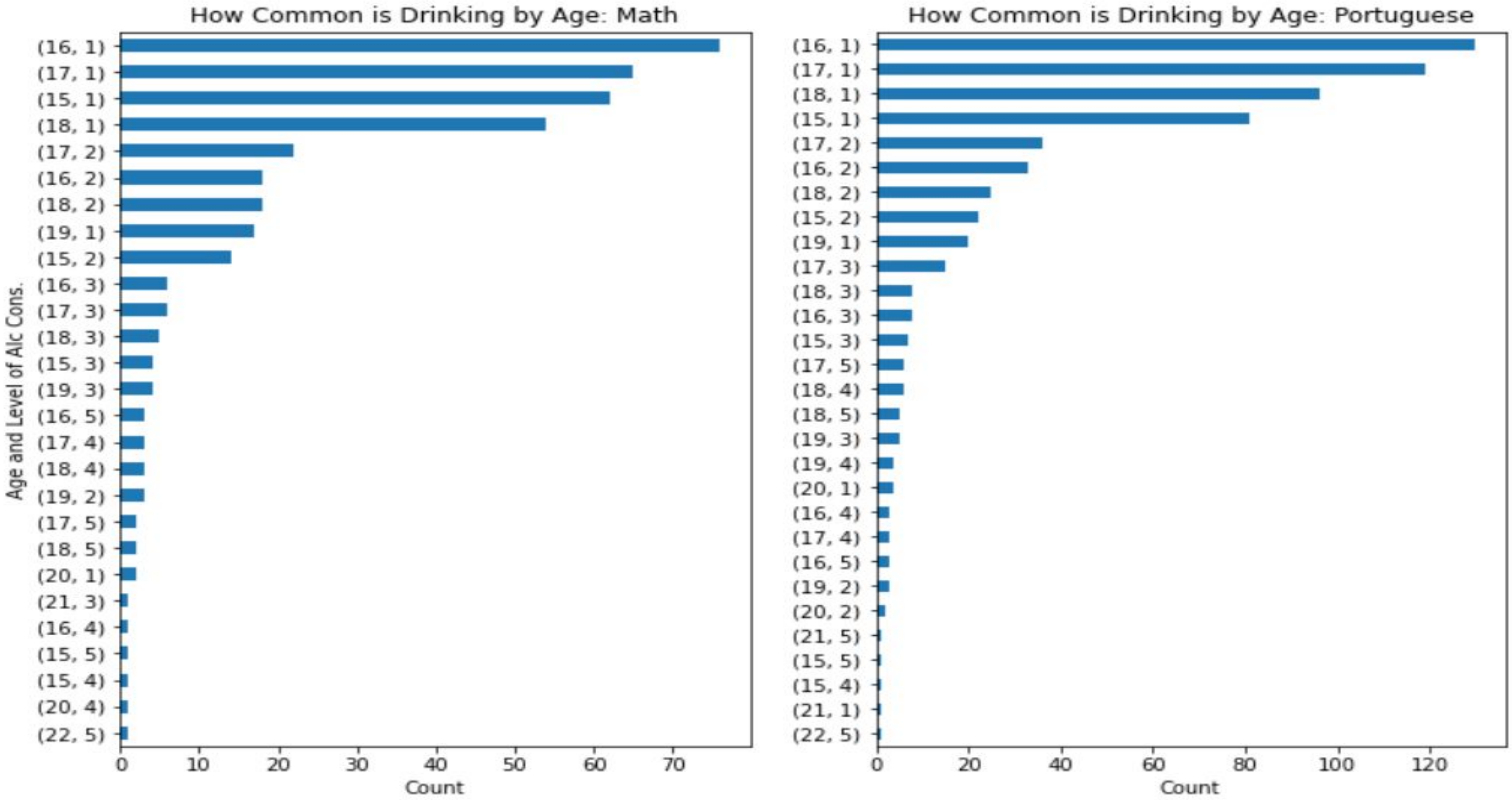
Visualizations

Methodology

- Libraries: Matplotlib.pyplot , and Altair
- Groupby [“independent variable”]
- Aggregate by the **mean** of G1, G2, G3
- Create multiple subplots using
  - fig, ax = plt.subplots(rows, cols)

‘G3’ Grade Distribution of Students

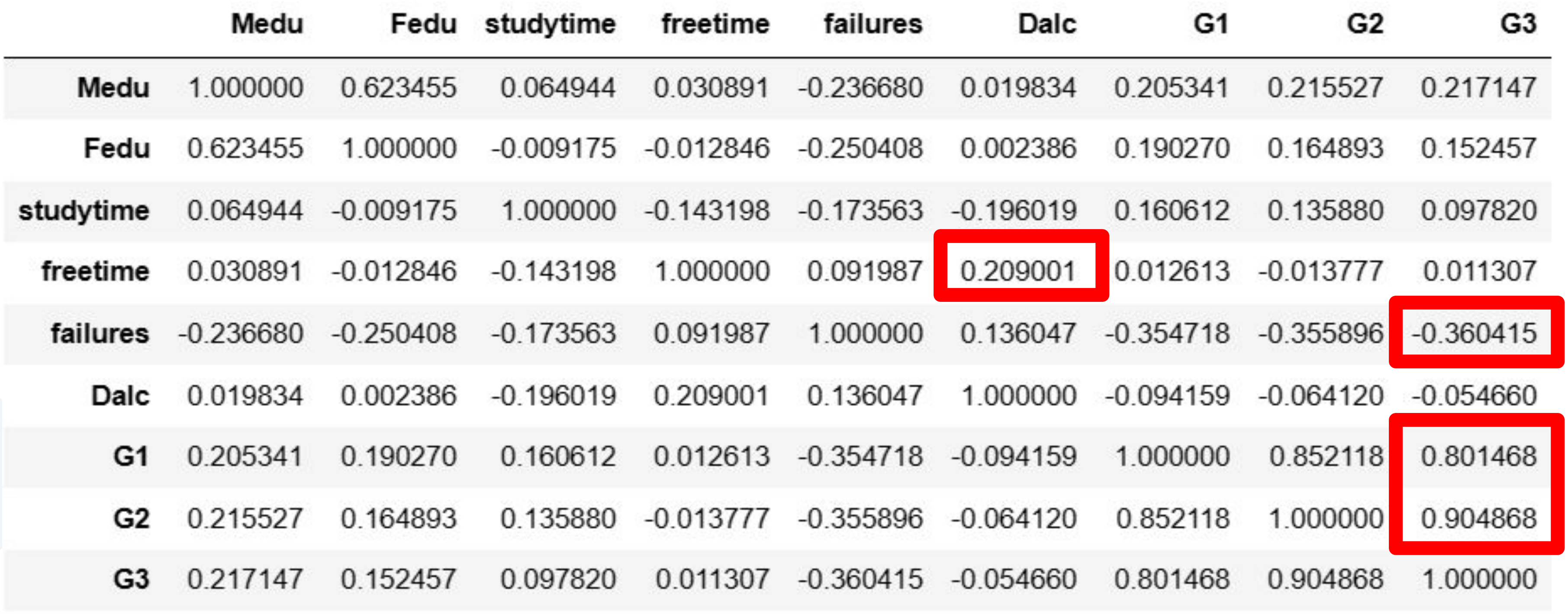




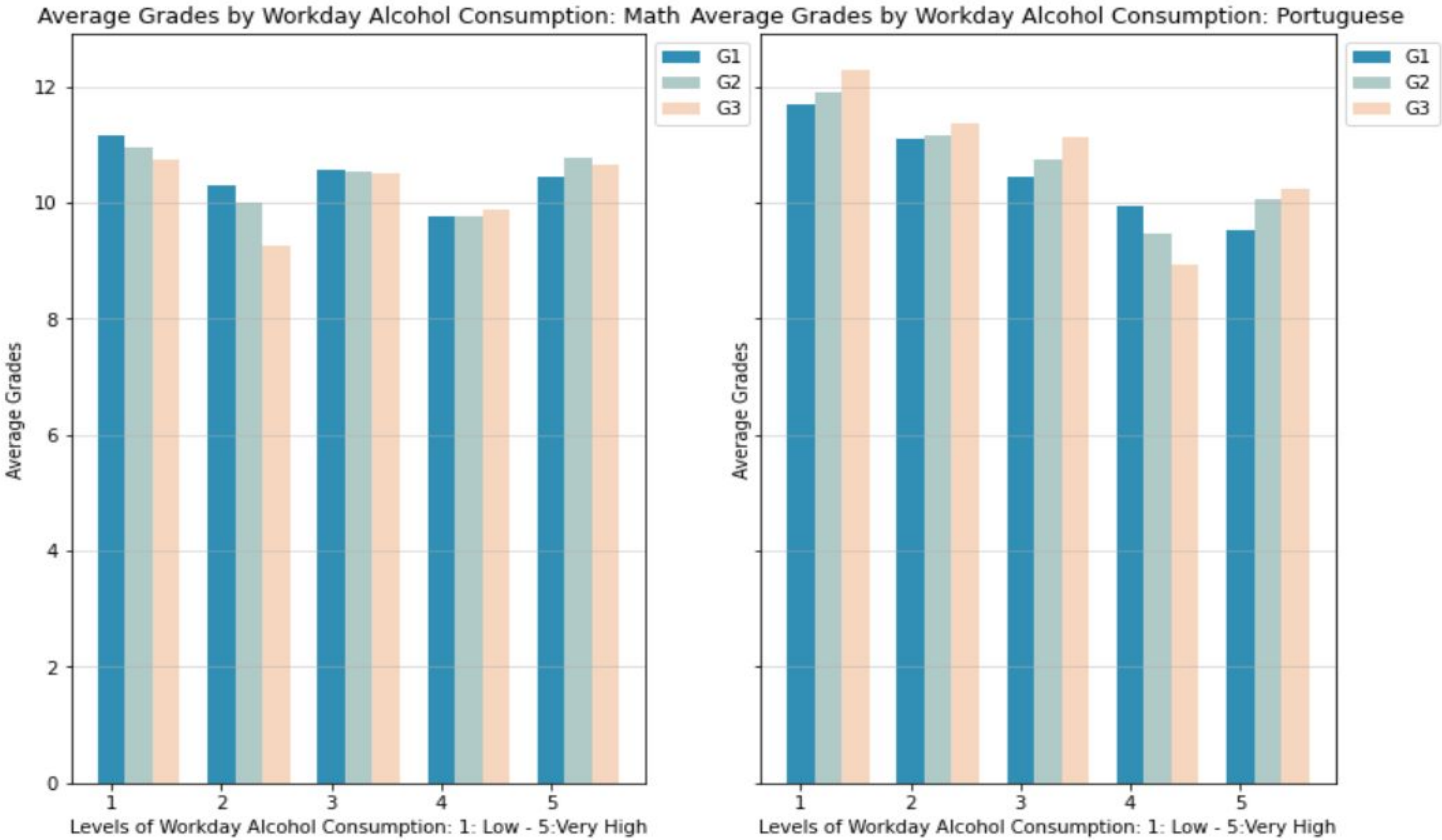
How Common is Drinking by Age: Math

How Common is Drinking by Age: Portuguese

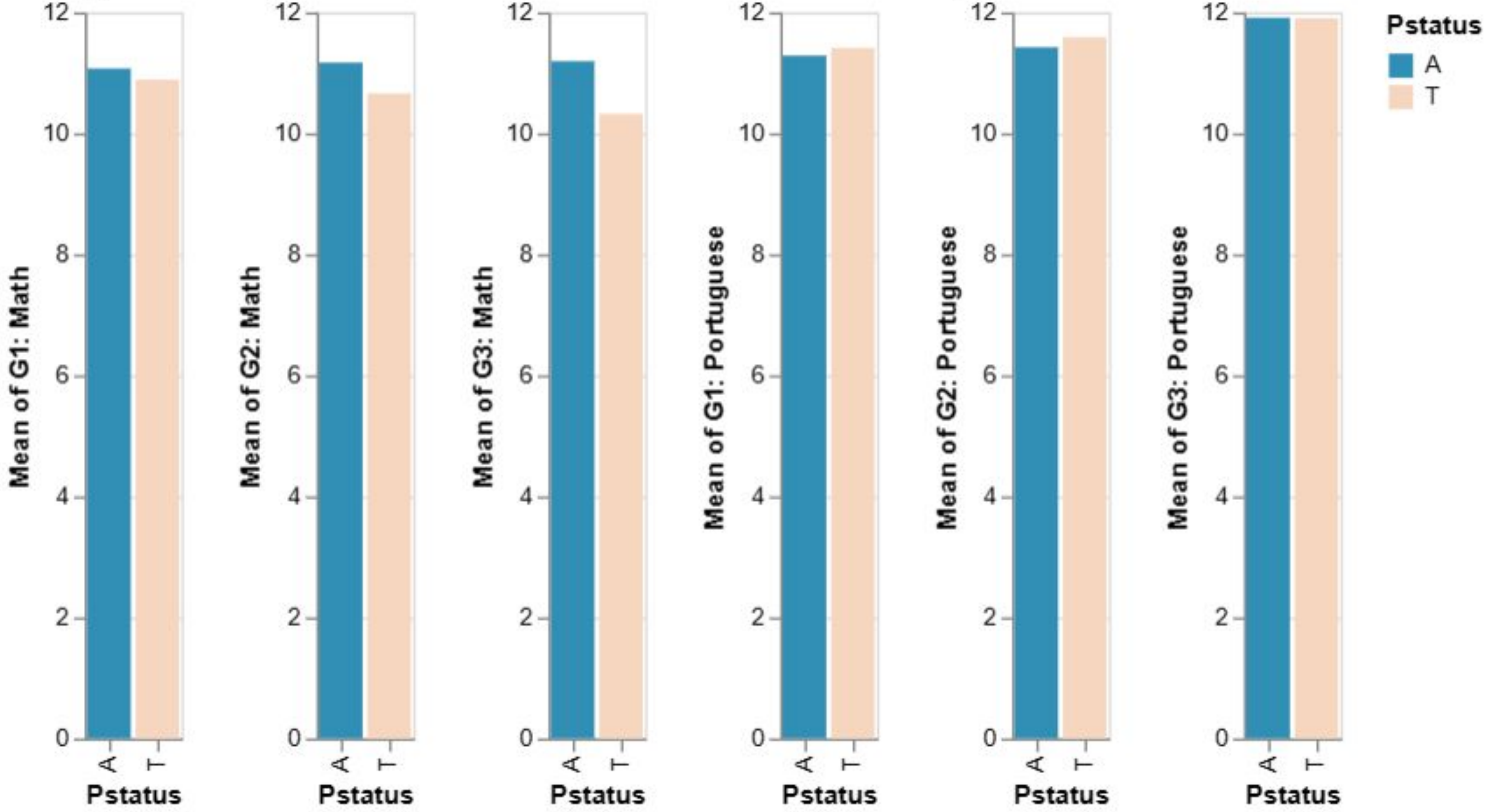
Correlation Between Variables

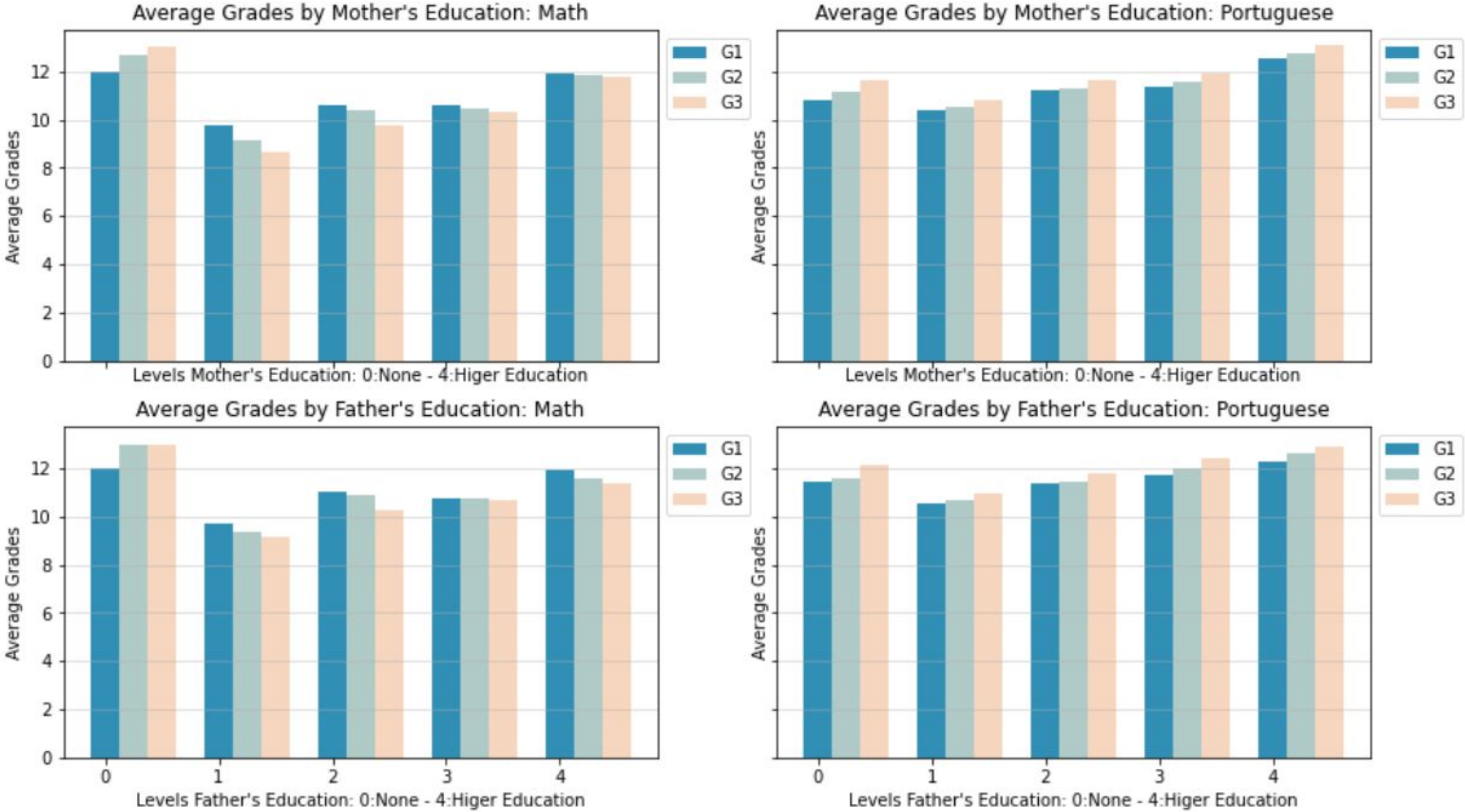


	Medu	Fedu	studytime	freetime	failures	Dalc	G1	G2	G3
Medu	1.000000	0.623455	0.064944	0.030891	-0.236680	0.019834	0.205341	0.215527	0.217147
Fedu	0.623455	1.000000	-0.009175	-0.012846	-0.250408	0.002386	0.190270	0.164893	0.152457
studytime	0.064944	-0.009175	1.000000	-0.143198	-0.173563	-0.196019	0.160612	0.135880	0.097820
freetime	0.030891	-0.012846	-0.143198	1.000000	0.091987	0.209001	0.012613	-0.013777	0.011307
failures	-0.236680	-0.250408	-0.173563	0.091987	1.000000	0.136047	-0.354718	-0.355896	-0.360415
Dalc	0.019834	0.002386	-0.196019	0.209001	0.136047	1.000000	-0.094159	-0.064120	-0.054660
G1	0.205341	0.190270	0.160612	0.012613	-0.354718	-0.094159	1.000000	0.852118	0.801468
G2	0.215527	0.164893	0.135880	-0.013777	-0.355896	-0.064120	0.852118	1.000000	0.904868
G3	0.217147	0.152457	0.097820	0.011307	-0.360415	-0.054660	0.801468	0.904868	1.000000


Average Grades by Workday Alcohol Consumption: Math
Average Grades by Workday Alcohol Consumption: Portuguese
G1
G2
G3
G1
G2
G3
Average Grades
Average Grades
0
2
4
6
8
10
12
1
2
3
4
5
1
2
3
4
5
Levels of Workday Alcohol Consumption: 1: Low - 5:Very High
Levels of Workday Alcohol Consumption: 1: Low - 5:Very High

Average Grades by Parent's Cohabitation Status: Math vs Portuguese





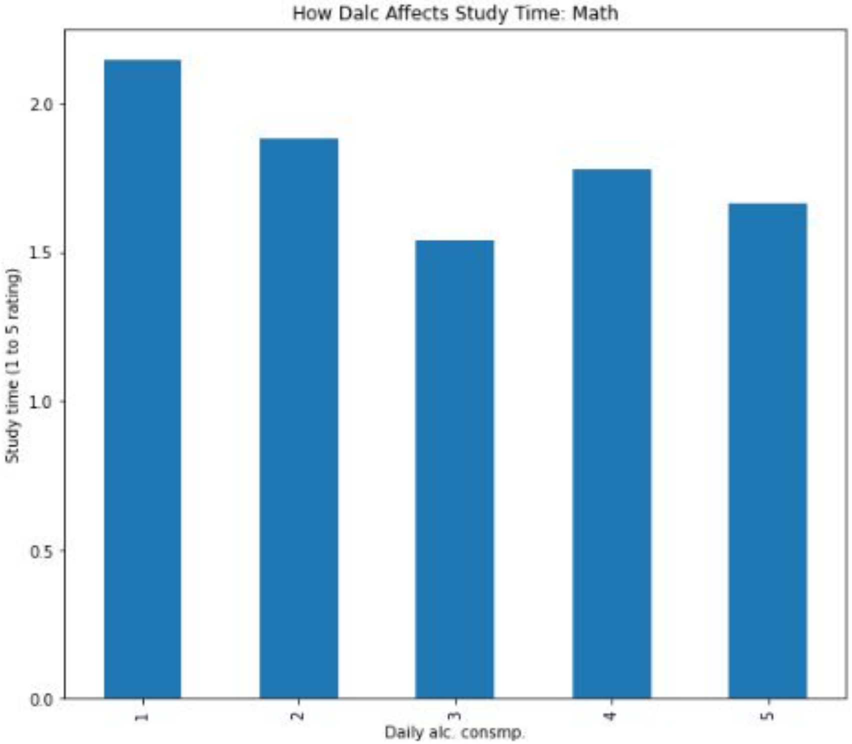
Average Grades by Mother's Education: Math

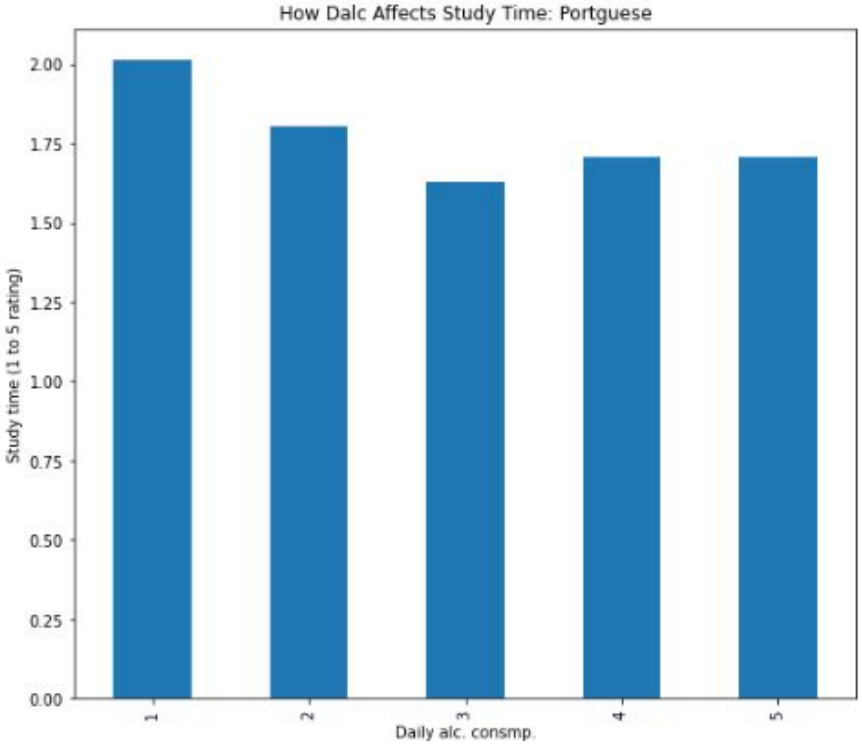
Average Grades by Mother's Education: Portuguese

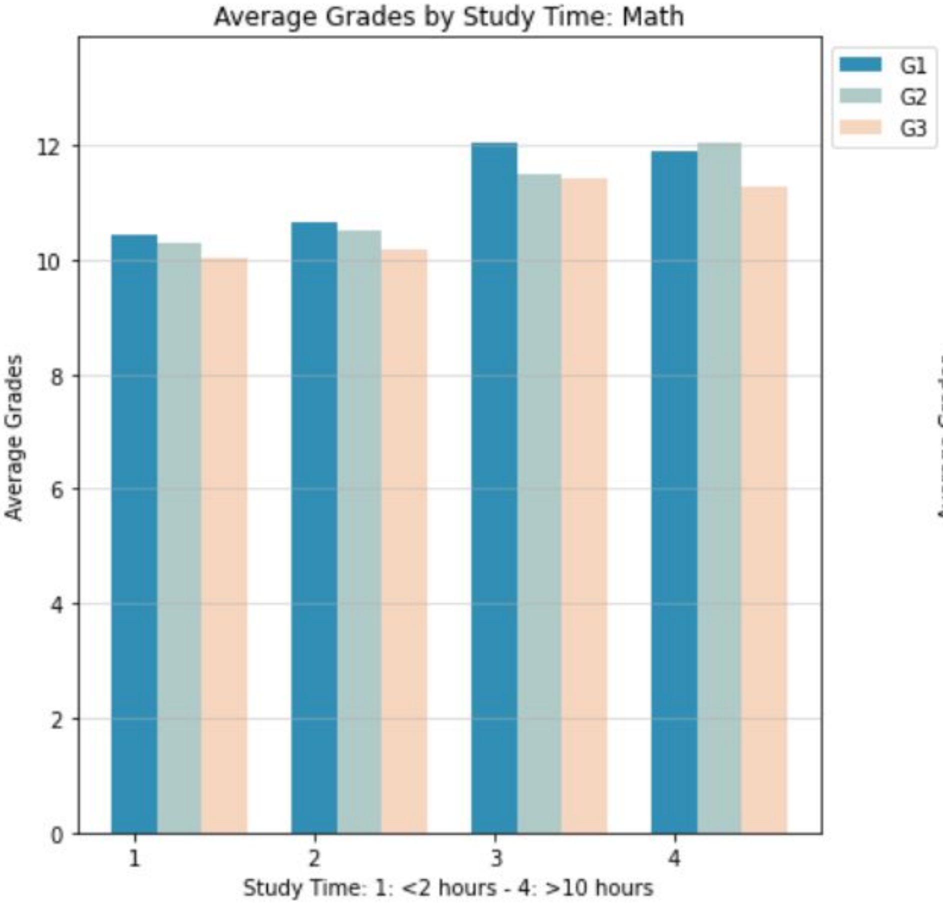
Average Grades by Father's Education: Math

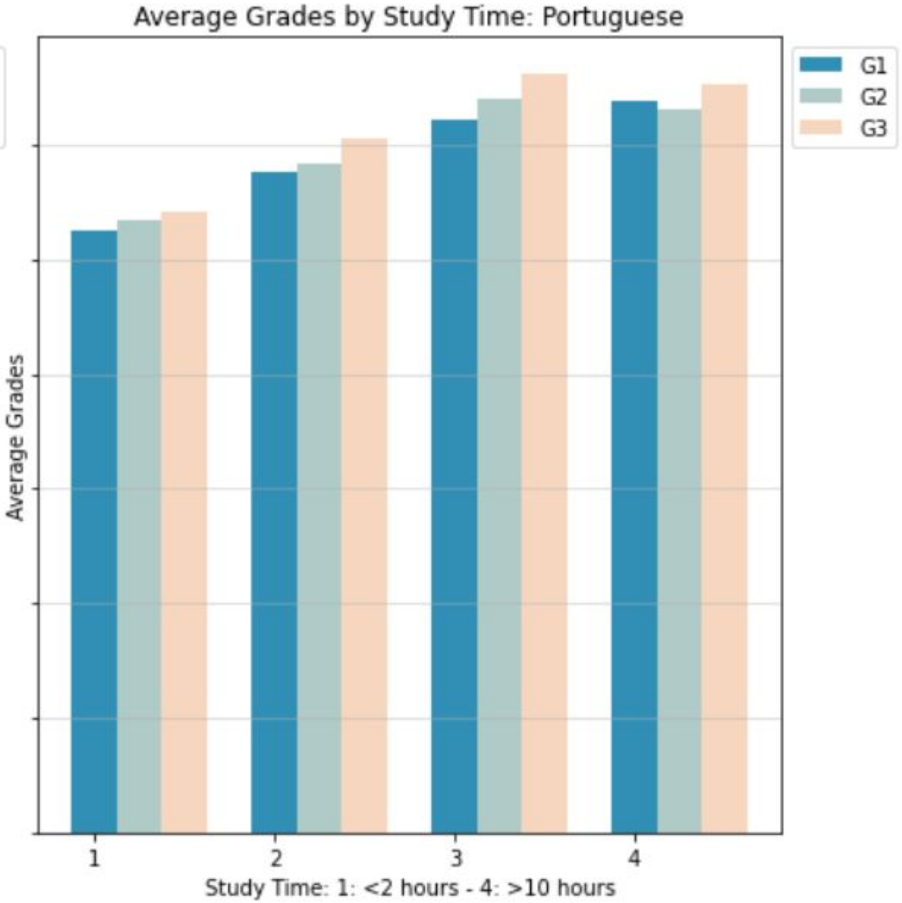
Average Grades by Father's Education: Portuguese

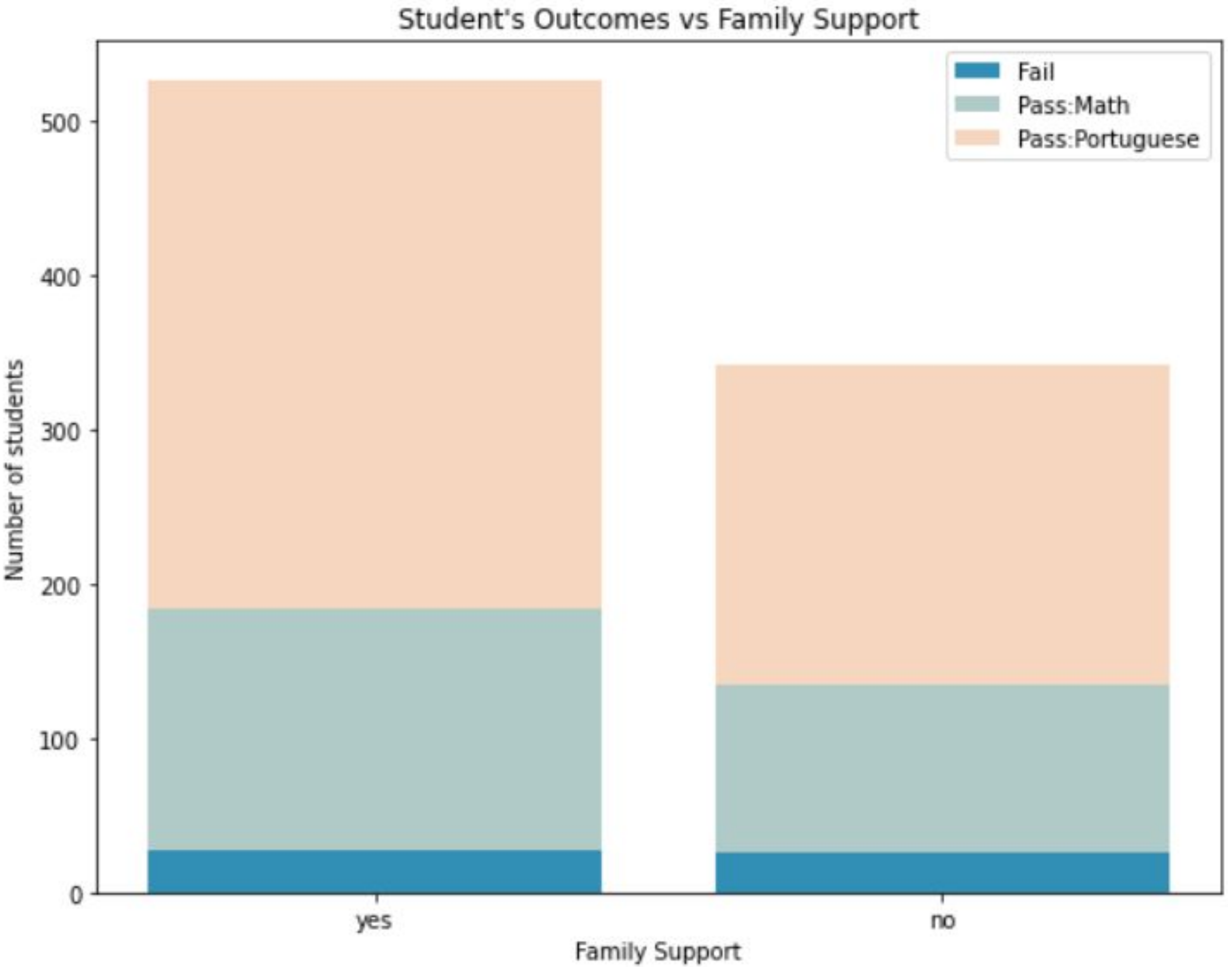
Include or not






Average Grades by Study Time: Math
Average Grades
0
2
4
6
8
10
12
1
2
3
4
Study Time: 1: <2 hours - 4: >10 hours
G1
G2
G3


Average Grades by Study Time: Portuguese
Average Grades
1
2
3
4
Study Time: 1: <2 hours - 4: >10 hours
G1
G2
G3



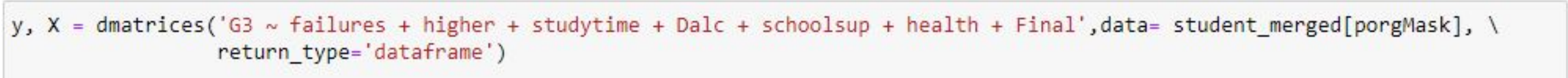
Proportion of family support - **‘No’** by Passing status:

Fail: 50.9%
Pass (math): 40.8%
Pass (Portuguese): 37.9%

03

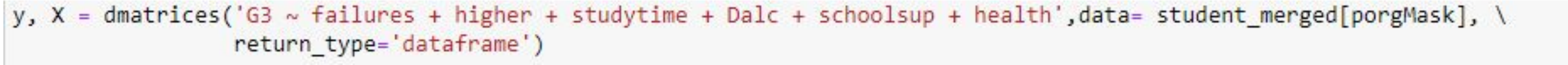
Regression/ Classification

Running a Regression Model

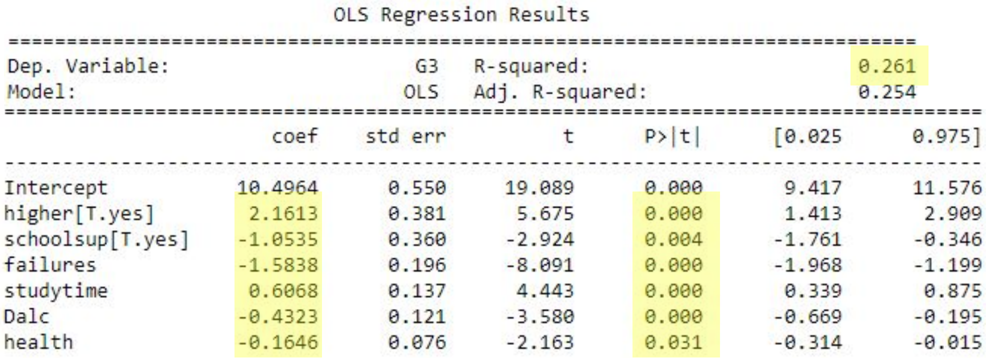


```
y, X = dmatrices('G3 ~ failures + higher + studytime + Dalc + schoolsup + health + Final',data= student_merged[porgMask], \
                 return_type='dataframe')
```

That doesn't make sense ('Final' attribute)!



```
y, X = dmatrices('G3 ~ failures + higher + studytime + Dalc + schoolsup + health',data= student_merged[porgMask], \
                 return_type='dataframe')
```



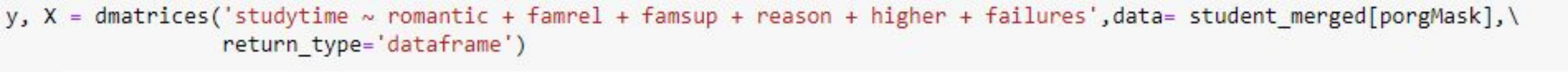
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     G3   R-squared:                       0.261
Model:                            OLS   Adj. R-squared:                  0.254
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept         10.4964      0.550     19.089      0.000       9.417      11.576
higher[T.yes]      2.1613      0.381      5.675      0.000       1.413       2.909
schoolsup[T.yes]  -1.0535      0.360     -2.924      0.004      -1.761      -0.346
failures          -1.5838      0.196     -8.091      0.000      -1.968      -1.199
studytime          0.6068      0.137      4.443      0.000       0.339       0.875
Dalc              -0.4323      0.121     -3.580      0.000      -0.669      -0.195
health            -0.1646      0.076     -2.163      0.031      -0.314      -0.015
```

Behind-the-scenes forward selection yielded higher, failures, and studytime as good predictors...

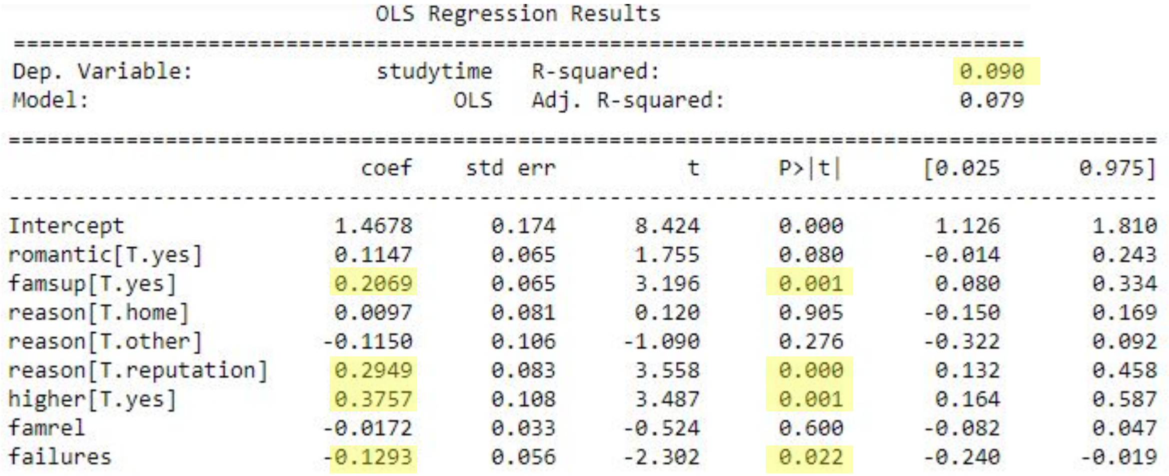
Caveat: R-squared

Another regression model…

Is studytime affected by students’ behaviors and upbringings?



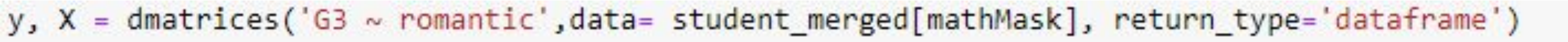
```
y, X = dmatrices('studytime ~ romantic + famrel + famsup + reason + higher + failures',data= student_merged[porgMask],\
                 return_type='dataframe')
```



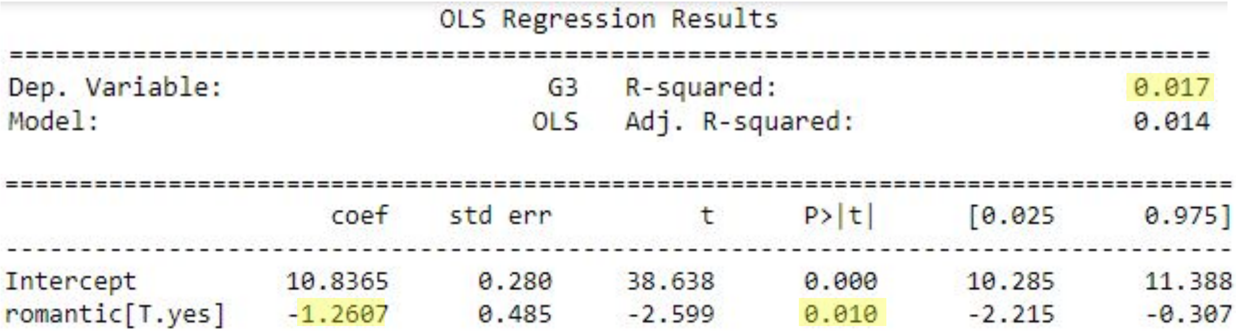
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              studytime   R-squared:                       0.090
Model:                            OLS   Adj. R-squared:                  0.079

==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                  1.4678      0.174      8.424      0.000       1.126       1.810
romantic[T.yes]            0.1147      0.065      1.755      0.080      -0.014       0.243
famsup[T.yes]              0.2069      0.065      3.196      0.001       0.080       0.334
reason[T.home]             0.0097      0.081      0.120      0.905      -0.150       0.169
reason[T.other]           -0.1150      0.106     -1.090      0.276      -0.322       0.092
reason[T.reputation]       0.2949      0.083      3.558      0.000       0.132       0.458
higher[T.yes]              0.3757      0.108      3.487      0.001       0.164       0.587
famrel                    -0.0172      0.033     -0.524      0.600      -0.082       0.047
failures                  -0.1293      0.056     -2.302      0.022      -0.240      -0.019
```

Yes, in some instances.

Interestingly…



```
y, X = dmatrices('G3 ~ romantic',data= student_merged[mathMask], return_type='dataframe')
```

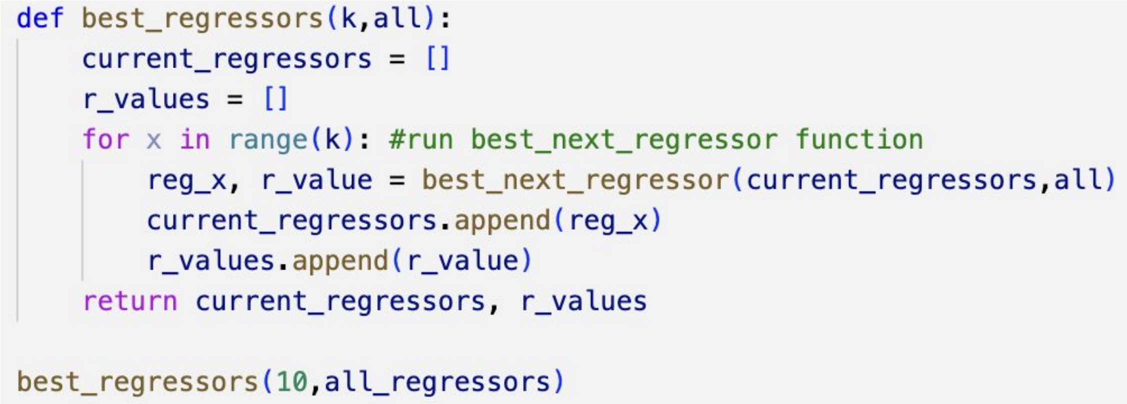


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     G3   R-squared:                       0.017
Model:                            OLS   Adj. R-squared:                  0.014

==================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          10.8365      0.280     38.638      0.000      10.285      11.388
romantic[T.yes]    -1.2607      0.485     -2.599      0.010      -2.215      -0.307
```

Takeaway: May not actually be indicative of any real world relationship . .

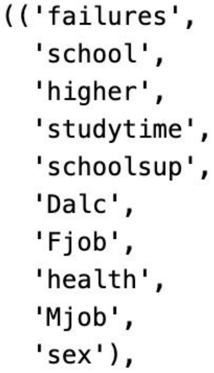
Feature Selection for Classification

Forward feature selection



```
def best_regressors(k,all):
    current_regressors = []
    r_values = []
    for x in range(k): #run best_next_regressor function
        reg_x, r_value = best_next_regressor(current_regressors,all)
        current_regressors.append(reg_x)
        r_values.append(r_value)
    return current_regressors, r_values

best_regressors(10,all_regressors)
```

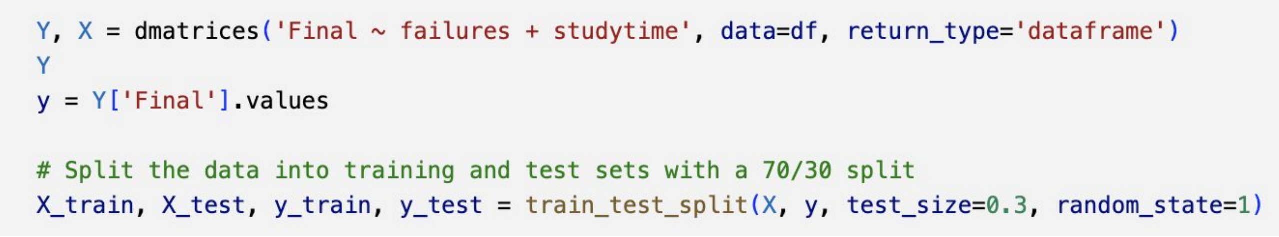
Top 10 features:



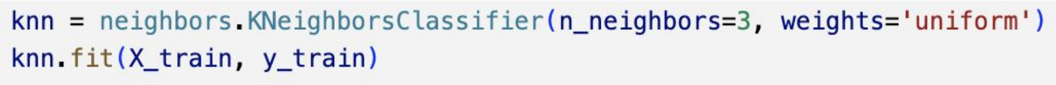
```
(('failures',
  'school',
  'higher',
  'studytime',
  'schoolsup',
  'Dalc',
  'Fjob',
  'health',
  'Mjob',
  'sex'),
```

Building a Classification Model (KNN)

What is a good combination of features?

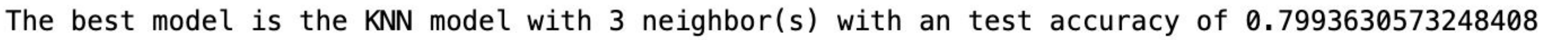


```
Y, X = dmatrices('Final ~ failures + studytime', data=df, return_type='dataframe')
Y
y = Y['Final'].values

# Split the data into training and test sets with a 70/30 split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```



```
knn = neighbors.KNeighborsClassifier(n_neighbors=3, weights='uniform')
knn.fit(X_train, y_train)
```

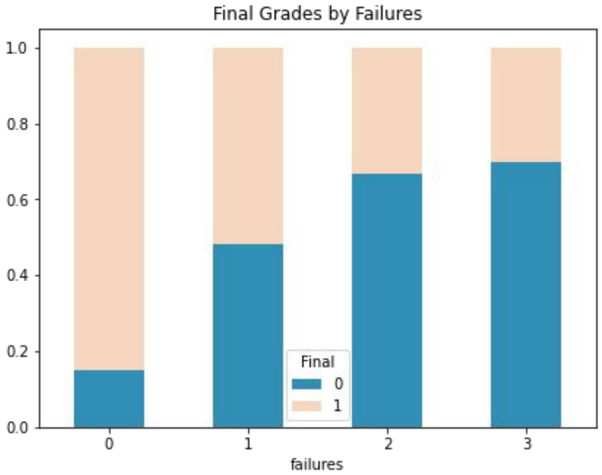
What is the best K?

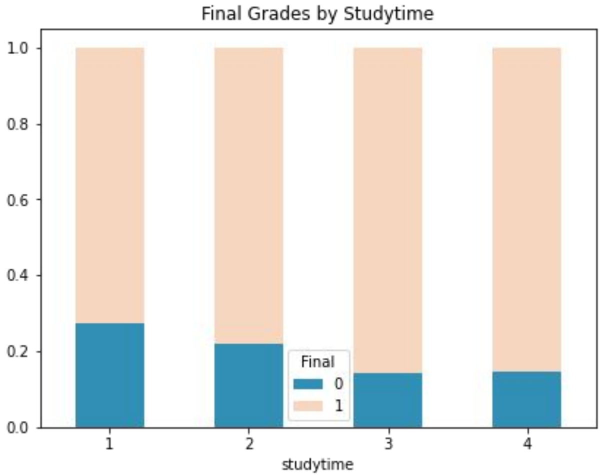


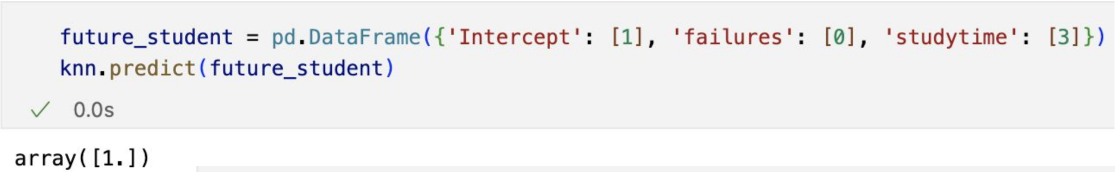
```
The best model is the KNN model with 3 neighbor(s) with an test accuracy of 0.7993630573248408
```

Building a Classification Model (KNN)

Do our predictors make sense?



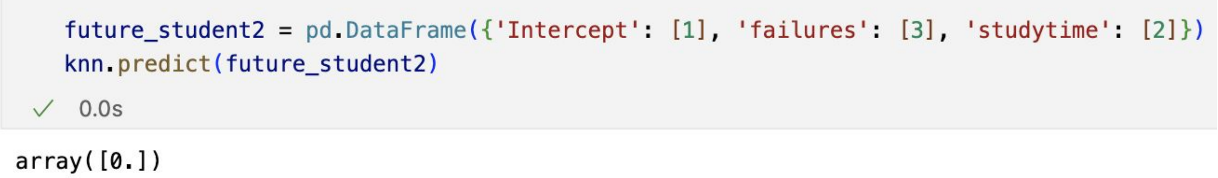




```
future_student = pd.DataFrame({'Intercept': [1], 'failures': [0], 'studytime': [3]})
knn.predict(future_student)
```

✓ 0.0s

```
array([1.])
```



```
future_student2 = pd.DataFrame({'Intercept': [1], 'failures': [3], 'studytime': [2]})
knn.predict(future_student2)
```

✓ 0.0s

```
array([0.])
```

An example use of this model…

04

Final Thoughts

Conclusions Going Forward

Conclusions

Data Takeaways

Effects on grades exist, may be weak

Caveats

Self-reported data may be inaccurate

Further Research

Larger sample, anonymized information