

Going the Distance: How Commute Time Affects Performance

Michael Govaerts

Abstract

A data set combination of student performance and potentially relevant student characteristics from the UCI Machine Learning Repository was used to help identify any relationship between commute time and student performance, which could potentially be extrapolated to influence commute-related educational decisions and public policy. The data sets were manipulated and analyzed using the pandas software library, written in the Python programming language, establishing clear correlations between data attributes. Matplotlib and seaborn, a matplotlib enhancement, were used to visualize the correlations between data set attributes, primarily through the use of pairplots. This report's analyses established a clear negative correlation between student commute time and performance from the data. However, due to the highly limited data set and statistical limitations of this report, additional research into commuting and human performance is needed to generate more comprehensive conclusions that might help influence public or educational policy.

Motivation

Long commutes are a tremendous problem for society to mitigate. More driving comes with significant hazards, from increased greenhouse gas emissions to a greater number of traffic accidents. As we determine what degree of resources we should allocate to addressing this issue, we need to quantify the relative impact commuting has on us and our surroundings. There are only so many resources we can invest in any of the many challenges our society is facing. Thus we must determine how detrimental commuting is before we can justify allocating limited resources to help mitigate it.

Additionally, as one of the critical suppliers of those public resources is tax revenue, determining how commute times affect our performance seems like a valuable area of study. Education is crucial for allowing individuals to reach their economic potential. Since many students may consider commuting longer distances to save money on housing costs, it seems valuable for them, their advocates, and their educational institutions to analyze whether or not the costs of their increased commute times outweigh the benefits.

To do this, we need a data set that can help determine the relationship between transit time and student performance. Fortunately, UC Irvine's Machine Learning Repository has a data set combination that can do just that.

Dataset(s)

This is a multivariate data set with integer attribute characteristics. There are 1044 data instances across 33 attributes in two data sets. There are no missing values. The data sets are kept at the UC Irvine Machine Learning Repository and were originally gathered from secondary school students in two Portuguese schools in Portugal's Alentejo region. The data were collected during the 2005-2006 school year. The data consist of various demographic and characteristic values for each student and describe their academic grades in each of two subjects at three different instances. The dataset combination includes an attribute key describing what each attribute name refers to and how the attribute was measured.

Specifically, the dataset was obtained from:

<https://archive.ics.uci.edu/ml/datasets/Student+Performance>

Data Preparation and Cleaning

Fortunately, the data set had no missing values, and the data was formatted appropriately for analysis through two csv data files.

The only issue was that pandas presumes that the values in a csv file are separated by commas for the `read_csv` function. However, a quick review of pandas documentation, and adjusting the code accordingly, allowed the data sets csv files with values separated by semicolons to be properly read by pandas.

Research Question(s)

When comparing time spent commuting to student performance measures, do any meaningful trends emerge when focusing on a sample of Portuguese secondary school students? Can these conclusions be extended to the wider population to help influence public policy?

Methods (1/3)

1) A Jupyter notebook was created for this project, which is included alongside this presentation. Pandas, a Python library, was imported into the notebook to analyze the data. Written in the Python programming language, pandas is a software library designed for data manipulation and analysis.

2) Pandas expects csv file data to be separated with commas by default. Thus, as the csv data sets used in this project were separated by semicolons instead, the `read_csv` function was modified to accommodate this and the data was successfully read with pandas.

3) The data set was then partially reviewed to confirm the csv was properly processed by pandas. An attribute key describing each abbreviated attribute name was linked in the project's Jupyter notebook to make the identities of the data set more accessible to a reviewer.

[The attribute key was originally given here, by UCI: <https://archive.ics.uci.edu/ml/datasets/Student+Performance#>]

4) The data set was then limited to values that seemed relevant to travel time and performance; namely the student's reason for selecting the school, where distance from home was a given reason for selection; study time for the students, which will be useful in determining how study time relates to performance; and student performance measures at three different course checkpoints. Number of absences were also included to see if they were correlated with travel time.

Methods (2/3)

5) Using pandas .describe(), the descriptive statistics of the values were then displayed and reviewed, where it was noted that very few students experienced significant travel time. This scarcity of data representing higher travel times may significantly reduce the conclusive power of this project.

6) Any initial trends from the descriptive statistics were explored further, in this case via pandas .mode() to confirm that the most frequent travel time value was the value representing the least travel time. It was also noted that the most frequent reason for selecting the school was not the school's distance from the student's home, which suggests that travel time was not highly important to most students.

7) Suspected correlations were then explored further with pandas .corr(), where notable correlative relationships were identified between the analyzed attributes. These relationships will be analyzed more in the Findings section of the project. Notably, it was identified that the number of absences was not significantly correlated with travel time.

8) To visualize the data, matplotlib was imported into the Jupyter notebook, along with seaborn, a matplotlib enhancement.

Methods (3/3)

9) A pairplot was then created using seaborn. Pairplot scatterplots allowed attributes to be visually compared against each other. Regression lines were also added to the pairplots using the kind="reg" parameter, allowing us to more easily visualize the correlations between attributes in each of the scatterplots. Additionally, histograms were formed when the pairplots caused an attribute to be compared against itself, allowing each attribute's relative frequencies to be compared in the attribute histograms.

Importantly, when the histograms are formed, the labeled y-axes of the pairplots are not accurate for the histograms, and so the y-axes for the histograms should largely be ignored and the height of the bars used instead to determine the relative frequency of these attribute values. [This will be much more apparent in Findings later, and is also evident in the included Jupyter notebook for the project.]

In our Jupyter notebook, sns.histplot() was then used for the travel time attribute, which was one of the histograms in the pairplot, to demonstrate how the true y-axis of the histograms is a measure of the frequency of the attribute values in a given discrete bin. This confirms that the y-axes listed in the seaborn pairplots should not be used for the histograms.

The scatterplot-and-histogram-visualized results will then be used to help identify and describe correlations and frequency distributions in our data set, which will contribute to our findings later in the report.

10) An image for the pairplot visualization was then created for use in reporting our findings.

11) The above procedure was then repeated for the csv file measuring attributes for the Portuguese class data, as this unique data helps provide more context to the trends we observe in the Math class data, allowing us to draw firmer conclusions from our data.

Findings (1/7)

As a result of our analysis, we were able to identify that the data demonstrated that the students' travel time to school was negatively correlated to their grades in the course. In the case of their math performance, the strength of the negative correlation for travel time was greater than the strength of the positive correlation for study time in terms of the student's final course grade, which should help underscore how powerfully negative commuting can be.

Thus, using the data sets in this project, we have been able to answer our research question and demonstrate that student travel time does appear to be related to student performance. In addition, we are able to confirm our suspicion that the relationship between commute and performance would be negative.

Findings (2/7)

We were able to easily identify these findings using the the correlation function on the attributes relevant to our analysis, namely:

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)

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)

Findings (3/7)

Math data set correlations:

	traveltime	studytime	G1	G2	G3
traveltime	1.000000	-0.100909	-0.093040	-0.153198	-0.117142
studytime	-0.100909	1.000000	0.160612	0.135880	0.097820
G1	-0.093040	0.160612	1.000000	0.852118	0.801468
G2	-0.153198	0.135880	0.852118	1.000000	0.904868
G3	-0.117142	0.097820	0.801468	0.904868	1.000000

Please note the negative correlations between travel time (“traveltime”) and the grade score at each of three grade checkpoints (“G1”, “G2”, and “G3”). In the case of the final two grades (“G2” and “G3”), the negative correlation of travel time is greater in magnitude than the associated positive correlation of study time.

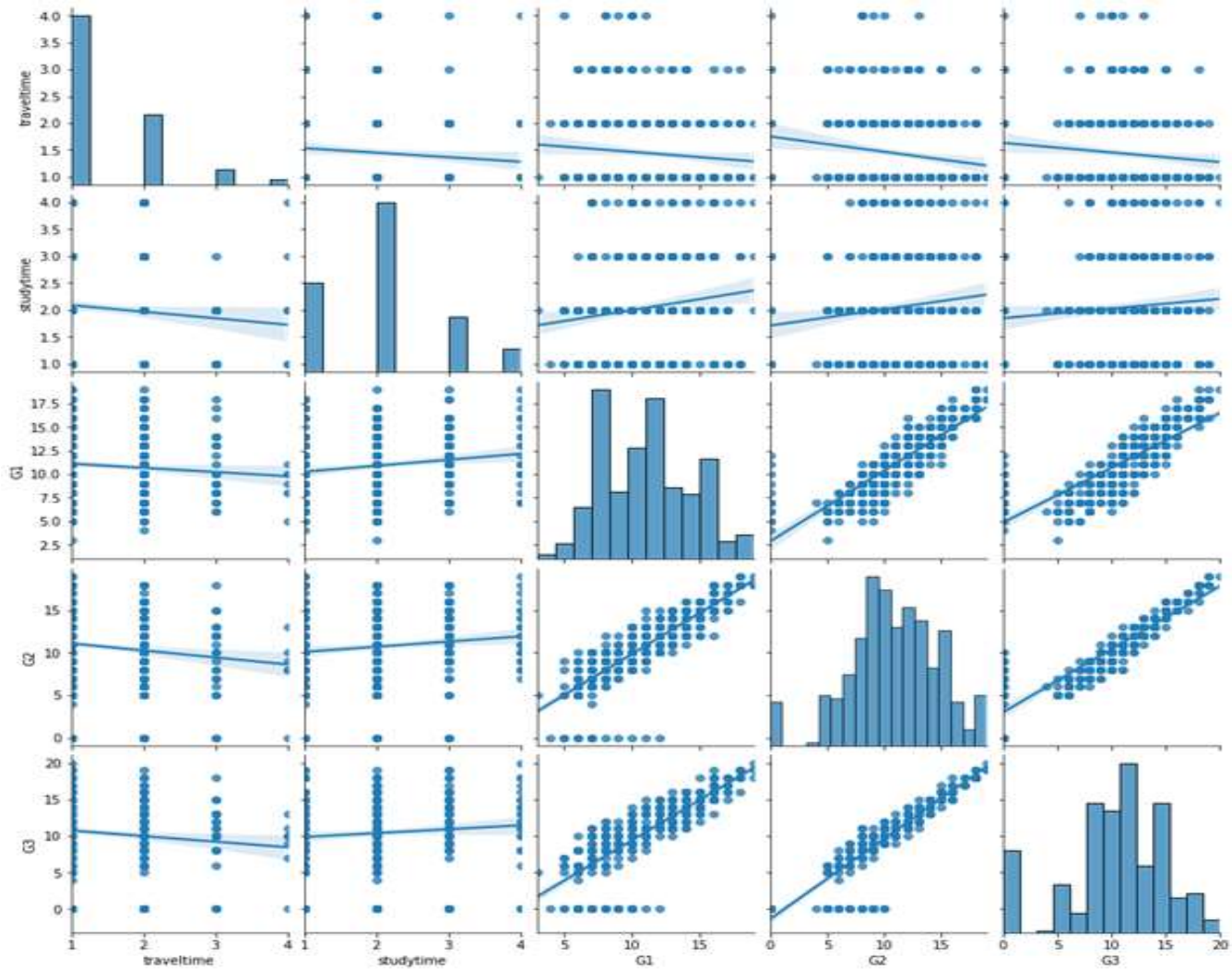
The pairplot on the next slide helps visualize these attribute relationships (in the form of scatterplots) and the frequency of the attribute scores (in the case of the histograms). [Please note, the y-axes of the visualizations only refer to the scatterplots. The y-axes of the histograms are just a general count of the frequency of a given attribute score. (The higher the histogram bar, the more frequently that score was reported.)]

Findings (4/7)

Math Data Set PairPlots (Scatterplots and Histograms)

Scatterplot regression lines help identify the correlation between the two attributes being compared.

(Notably, the scatterplots in the upper right show a negative correlation between travel time and the scores at each grade checkpoint.)



Findings (5/7)

Portuguese data set correlations

	traveltime	studytime	G1	G2	G3
traveltime	1.000000	-0.063154	-0.154120	-0.154489	-0.127173
studytime	-0.063154	1.000000	0.260875	0.240498	0.249789
G1	-0.154120	0.260875	1.000000	0.864982	0.826387
G2	-0.154489	0.240498	0.864982	1.000000	0.918548
G3	-0.127173	0.249789	0.826387	0.918548	1.000000

As we can see from the correlations above, travel time is again negatively correlated with performance at each of the grade checkpoints.

Additionally, in the next slide, we can see how the pairplots help illustrate these correlations.

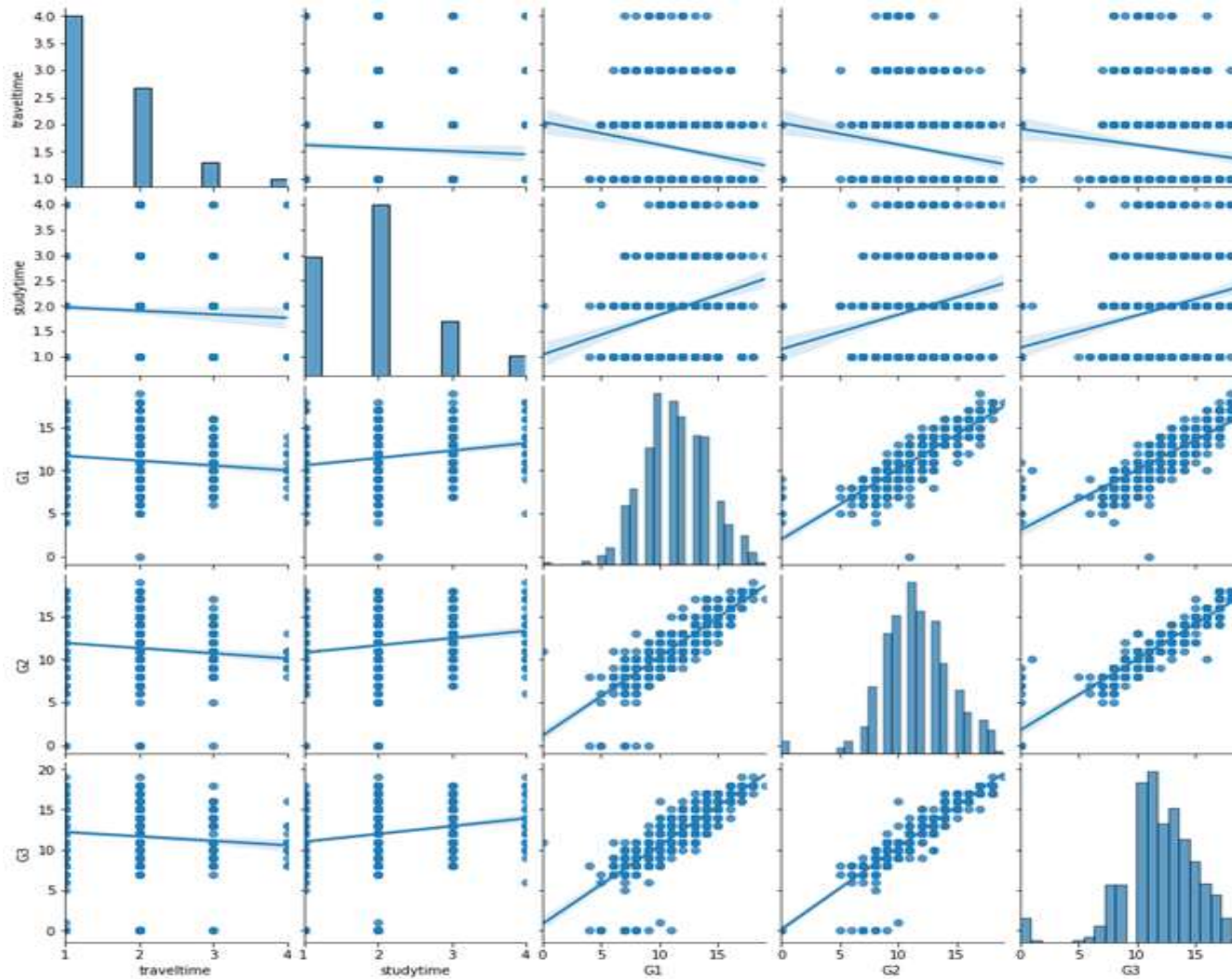
Findings (6/7)

Portuguese Data Set PairPlots

- Scatterplot regression lines help identify the correlation between the two attributes being compared.

- (Notably, the scatterplots in the upper right show a negative correlation between travel time and the scores at each grade checkpoint.

- Visually, we can see that these negative correlations between scores and travel time are more negative for Portuguese scores than for the math scores in the previous data set.)



Findings 7/7

As we can see from the visualizations, the scatterplots' regression lines identify the correlations between the two attributes being compared in each scatterplot.

In the case of both Portuguese and Math courses for these Portuguese students, travel time is negatively correlated with course grades, strongly suggesting that travel time impacts student performance, which is the question we were hoping to answer in our analysis. We were also able to better assess how travel time influenced student performance relative to other factors, like study time. Unfortunately, due to the limited data sets, it won't be easy to extend these conclusions to the broader community or influence public policy without more robust supporting evidence.

The histograms of these visualizations are also valuable, as they help us better identify the frequency of an attribute's scores. Importantly for our analyses, the travel time histograms in the upper left of the pairplots identify that most students in both course data sets have travel time attribute scores reflecting the least amount of travel time. This lack of high travel time data is an unfortunate limitation on our ability to analyze the effect of high commute time on student performance.

We can also identify other findings from the scatterplots and histograms that are less relevant to our research question, such as the visual suggestion that not all students took all the exams based on their resulting scores of zero. These outlier scores represent another limitation of our study that may affect our ability to draw meaningful conclusions from the data.

Limitations (1/2)

Unfortunately, while the data set was very accessible and did not require a significant amount of cleaning for use in our project, some clear limitations prevent us from drawing firm conclusions about how commute time may affect performance.

We must recall that the data set contained data from the 2005-2006 school year. Students were from the Alentejo area of Portugal, attending one of two public schools. This highly specific data set means that it will be challenging to extend the conclusions from our analyses very far. Also, the fact that we only have 1044 records across both data sets suggests the conclusions we can draw are much less powerful than if we had a much larger number of records to work with, especially a larger number of records from a more diverse sample population.

Additionally, as our statistical proficiency is not very developed based on this course's scope, there is likely more we could have done using more advanced statistical methods to improve on our analyses and the resultant conclusions. Another clear limitation based on this lack of statistical prowess is the potential for one of the other attributes to confound the relationship between travel time and performance.

Future research could therefore involve applying more advanced statistical methods to determine if more conclusive results could be obtained from these data sets. However, based on the highly specific nature of the data set, the results may not be particularly meaningful to our wider understanding of commute time and performance.

Limitations (2/2)

Additionally, as seen in our Jupyter notebook for the project, absences were notably negatively correlated with grades at each checkpoint. This negative correlation was similar to the negative correlation between travel time and checkpoint grades.

However, absences were only very weakly negatively correlated with travel time. Thus, to draw more firm conclusions about how travel time, checkpoint grades, and other attributes may be related, more elaborate statistical methods are likely needed. It would seem that establishing these more complicated relationships would be necessary to remove their influence on the critical relationship we are looking to study: commute time and performance.

Thus, while this research question represents a valuable starting point, there is obviously much more that would need to be done in order to draw more compelling conclusions that may, in turn, influence public policy and government spending.

Conclusions (1/2)

Through this analysis, we were able to identify that time spent commuting (“traveltime”) had a measurable and consistently negative relationship with student performance in Math and Portuguese courses in our two-school sample of Alentejo-area Portuguese secondary students from the 2005-2006 school year. In fact in the Math course, school commute time appeared to have a stronger relationship with student performance at two of the three grade checkpoints than time spent studying did. However, in both courses, while time spent studying was correlated with higher grades, time spent commuting was associated with lower grades.

While we would likely need to use more advanced statistical methods and use a larger sample set to draw more decisive conclusions, it seems that we have clearly identified a trend in school commute’s relationship with student performance in our sample population. This trend also helped confirm the suspicion that led to our research question: that commute time would be associated with adverse outcomes in student performance.

Conclusions (2/2)

Unfortunately, as mentioned, without a more comprehensive statistical analysis and a much larger data set, it would be difficult to justify extending our conclusions to larger populations, especially in designating public funding. However, our project should help justify the need for additional research into the harmful effects of commute time and how we might work to mitigate those effects.

So, while we might not have definitively answered the research question for society, we hope that this research has helped contribute to the overall discussion of commuting time and performance.

Acknowledgments

The author would like to thank the UC Irvine Machine Learning Repository for hosting the data sets used in this analysis. In addition, we would like to thank the Portuguese researchers who helped obtain the data and consented to have it hosted at the Machine Learning Repository.

Also, we are incredibly grateful to the scientists, especially data scientists, who are kind enough to share their knowledge and expertise with the rest of us. Reviewing the publicly available work of various data scientists, and observing their online discussions, significantly helped inform the direction of this project.

Thank you, as well, to the instructors and participants in this course for allowing the author to become more proficient in data science.

References

UC Irvine Machine Learning Repository:

<https://archive.ics.uci.edu/ml/datasets/Student+Performance>

Study generating this data set:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUTURE BUSINESS TECHNOLOGY Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[<http://www3.dsi.uminho.pt/pcortez/student.pdf>]

Going the Distance

Author: Michael Govaerts

```
In [352]: import pandas as pd
```

```
In [353]: # Read csv file, noting that the data is separated with semicolons rather than the default  
df = pd.read_csv("student(demo+grades)\student-mat.csv", sep=';')
```

Attribute Information:

Source:

<http://web.archive.org/web/20210209231731/https://archive.ics.uci.edu/ml/datasets/Student+Perform>

#

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

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

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

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

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

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

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

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

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

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

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

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

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

13 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)

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

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

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

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

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

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

- 20 nursery - attended nursery school (binary: yes or no)
- 21 higher - wants to take higher education (binary: yes or no)
- 22 internet - Internet access at home (binary: yes or no)
- 23 romantic - with a romantic relationship (binary: yes or no)
- 24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- 25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- 26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- 27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 29 health - current health status (numeric: from 1 - very bad to 5 - very good)
- 30 absences - number of school absences (numeric: from 0 to 93)

#

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

- 31 G1 - first period grade (numeric: from 0 to 20)
- 31 G2 - second period grade (numeric: from 0 to 20)
- 32 G3 - final grade (numeric: from 0 to 20, output target)

```
In [355]: #Review data set in tabular form to confirm it has been processed correctly

df.head()
```

```
Out[355]:
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	travel
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	

```
In [356]: # Limit data set to values that seem relevant to travel time and performance

df2 = df[['reason', 'traveltime', 'studytime', 'activities', 'G1', 'G2', 'G3', 'absences']
```

```
In [357]: df2.head()
```

```
Out[357]:
```

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course	2	2	no	5	6	6	6
1	course	1	2	no	5	5	6	4
2	other	1	2	no	7	8	10	10
3	home	1	3	yes	15	14	15	2

	reason	traveltime	studytime	activities	G1	G2	G3	absences
4	home	1	2	no	6	10	10	4

```
In [358]: # Review descriptive statistics of values to begin to analyze data
df2.describe()
```

```
Out[358]:
```

	traveltime	studytime	G1	G2	G3	absences
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	1.448101	2.035443	10.908861	10.713924	10.415190	5.708861
std	0.697505	0.839240	3.319195	3.761505	4.581443	8.003096
min	1.000000	1.000000	3.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	8.000000	9.000000	8.000000	0.000000
50%	1.000000	2.000000	11.000000	11.000000	11.000000	4.000000
75%	2.000000	2.000000	13.000000	13.000000	14.000000	8.000000
max	4.000000	4.000000	19.000000	19.000000	20.000000	75.000000

Unfortunately, we are able to see from this data set that few students traveled very far to get to the school they attended.

```
In [359]: # Confirm trend away from long travel times to school. As we can see from the below outp
df2.mode()
```

```
Out[359]:
```

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course	1	2	yes	10	9	10	0

```
In [360]: # Determine if there are any clear correlations between travel time, study time, absences
df2.corr()
```

```
Out[360]:
```

	traveltime	studytime	G1	G2	G3	absences
traveltime	1.000000	-0.100909	-0.093040	-0.153198	-0.117142	-0.012944
studytime	-0.100909	1.000000	0.160612	0.135880	0.097820	-0.062700
G1	-0.093040	0.160612	1.000000	0.852118	0.801468	-0.031003
G2	-0.153198	0.135880	0.852118	1.000000	0.904868	-0.031777
G3	-0.117142	0.097820	0.801468	0.904868	1.000000	0.034247
absences	-0.012944	-0.062700	-0.031003	-0.031777	0.034247	1.000000

As we can see from the table above, travel time is negatively correlated with study time and course grades at all periods.

Interestingly, travel time appears to have a stronger negative correlation with final course grade than the positive correlation that study time has with final course grade. This outcome may suggest that

travel time could be a larger factor in student performance than time spent studying. The only time this appears to not have been the case is for the first grade report. Absences are also not very strongly correlated with any of the other performance characteristics, which is surprising.

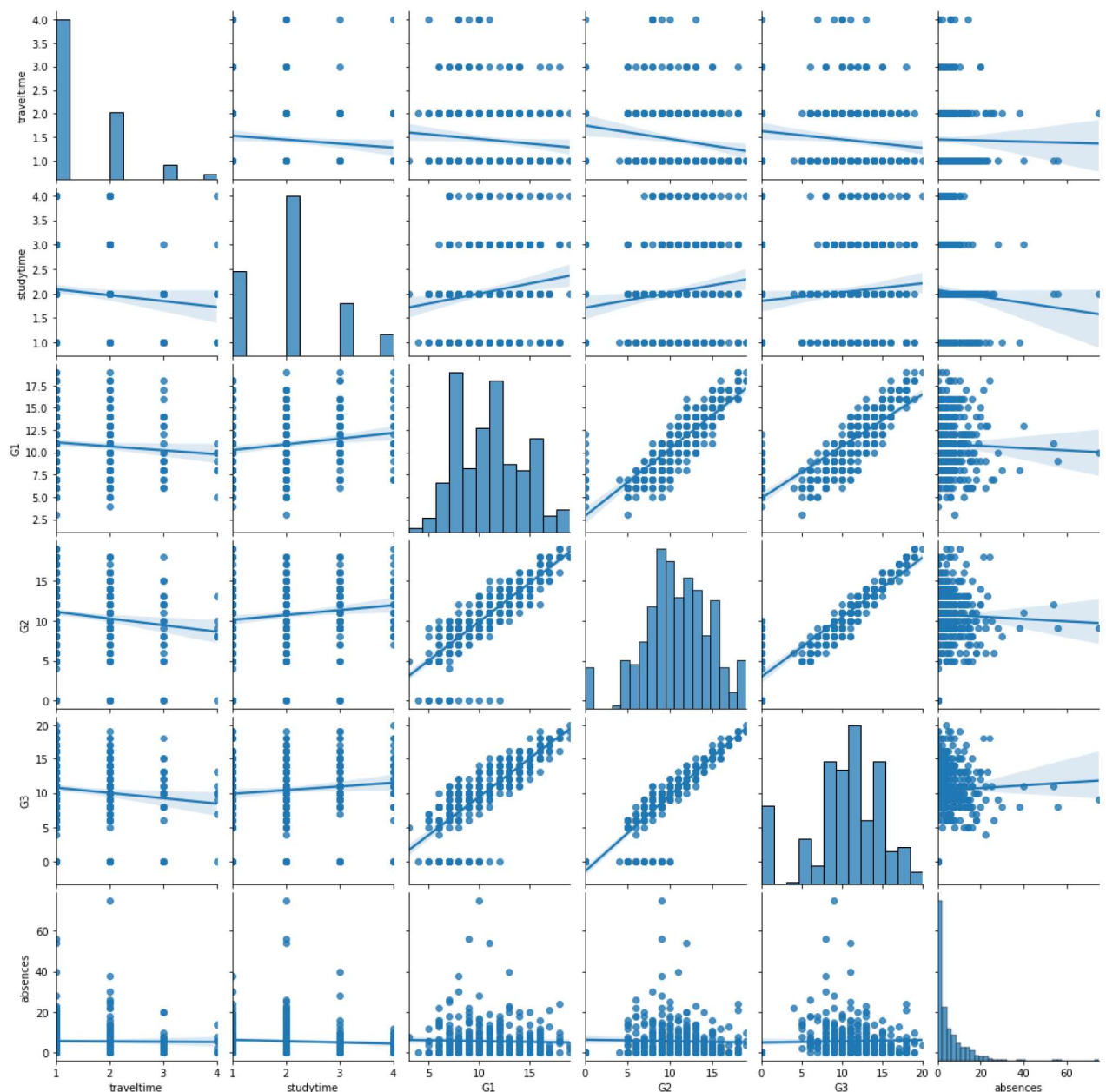
```
In [361]: # Matplotlib, a plotting library for Python, was imported for visualization purposes

import matplotlib.pyplot as plt

# Seaborn, a matplotlib enhancement, was imported to aid in visualization

import seaborn as sns
```

```
In [362]: # A pairplot was drawn using seaborn scatterplots
# The kind="reg" parameter was used to add linear regression models to the scatter plots,
# the visualization of any correlations
g = sns.pairplot(df2, kind="reg")
```

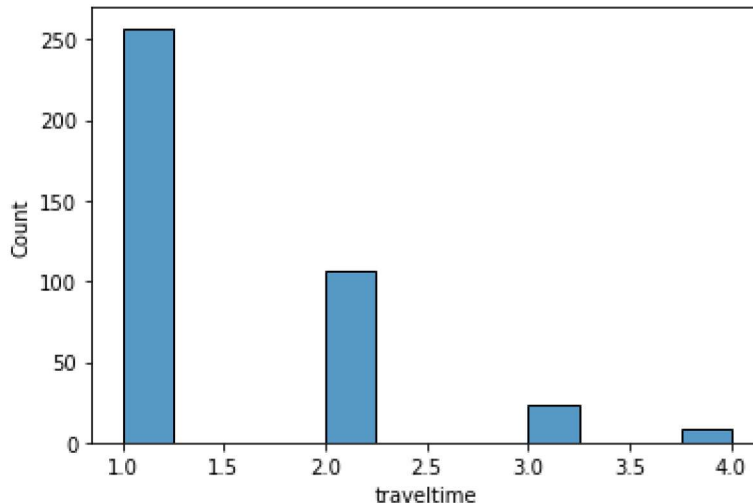


```
In [363]: # Save pairplot for use in our project report
```

```
g.savefig("math_output.png")
```

```
In [364]: # The axes labels of a seaborn pairplot, shown above, correspond to the relational scatter
# not the histograms that occur on the diagonal of the pairplot grid. However, as we can
# plot below, the heights of the histogram bars correspond to the relative frequency of e
# above.

h = sns.histplot(df2['traveltime']);
```



Unfortunately, both the pairplot and the individual traveltime histogram show very few students have significant travel time, so it would be hard to draw firm conclusions about the effects of long travel times using the existing data set.

Now let's see if the same trend exists for the Portuguese grades for the students.

```
In [365]: # Read csv file, noting that the data is separated with semicolons rather than the default

df1 = pd.read_csv("student(demo+grades)\student-por.csv", sep=';')
```

Attribute Information:

Source:

<http://web.archive.org/web/20210209231731/https://archive.ics.uci.edu/ml/datasets/Student+Perform>

#

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

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

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

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

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

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

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

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

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

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

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

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

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

13 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)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

#

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

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

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

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

In [367]:

```
#Review data set in tabular form to confirm it has been processed correctly

df1.head()
```

Out[367]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	travel
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	

```
In [368]: # Limit data set to values that seem relevant to travel time and performance

df1a = df1[['reason', 'traveltime', 'studytime', 'activities', 'G1', 'G2', 'G3', 'absences']]
```

```
In [369]: df1a.head()
```

```
Out[369]:
```

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course		2	no	0	11	11	4
1	course		1	no	9	11	11	2
2	other		1	no	12	13	12	6
3	home		1	yes	14	14	14	0
4	home		1	no	11	13	13	0

```
In [370]: # Review descriptive statistics of values to begin to analyze data

df1a.describe()
```

```
Out[370]:
```

	traveltime	studytime	G1	G2	G3	absences
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	1.568567	1.930663	11.399076	11.570108	11.906009	3.659476
std	0.748660	0.829510	2.745265	2.913639	3.230656	4.640759
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	10.000000	10.000000	10.000000	0.000000
50%	1.000000	2.000000	11.000000	11.000000	12.000000	2.000000
75%	2.000000	2.000000	13.000000	13.000000	14.000000	6.000000
max	4.000000	4.000000	19.000000	19.000000	19.000000	32.000000

Unfortunately, again, we are able to see from this data set that few students traveled very far to get to the school they attended.

We can also see that the 50%-ile student scores for Portuguese and Math were fairly similar across the data sets.

Interestingly, we have 649 respondents for the Portuguese grade data set, whereas we only had 395 for the Math grade data set.


```
In [371]: # Confirm trend away from long travel times to school. As we can see from the below output
# minimal travel time.

df1a.mode()
```

```
Out[371]:
```

	reason	traveltime	studytime	activities	G1	G2	G3	absences
0	course	1	2	no	10	11	11	0

```
In [372]: # Determine if there are any clear correlations between travel time, study time, absences

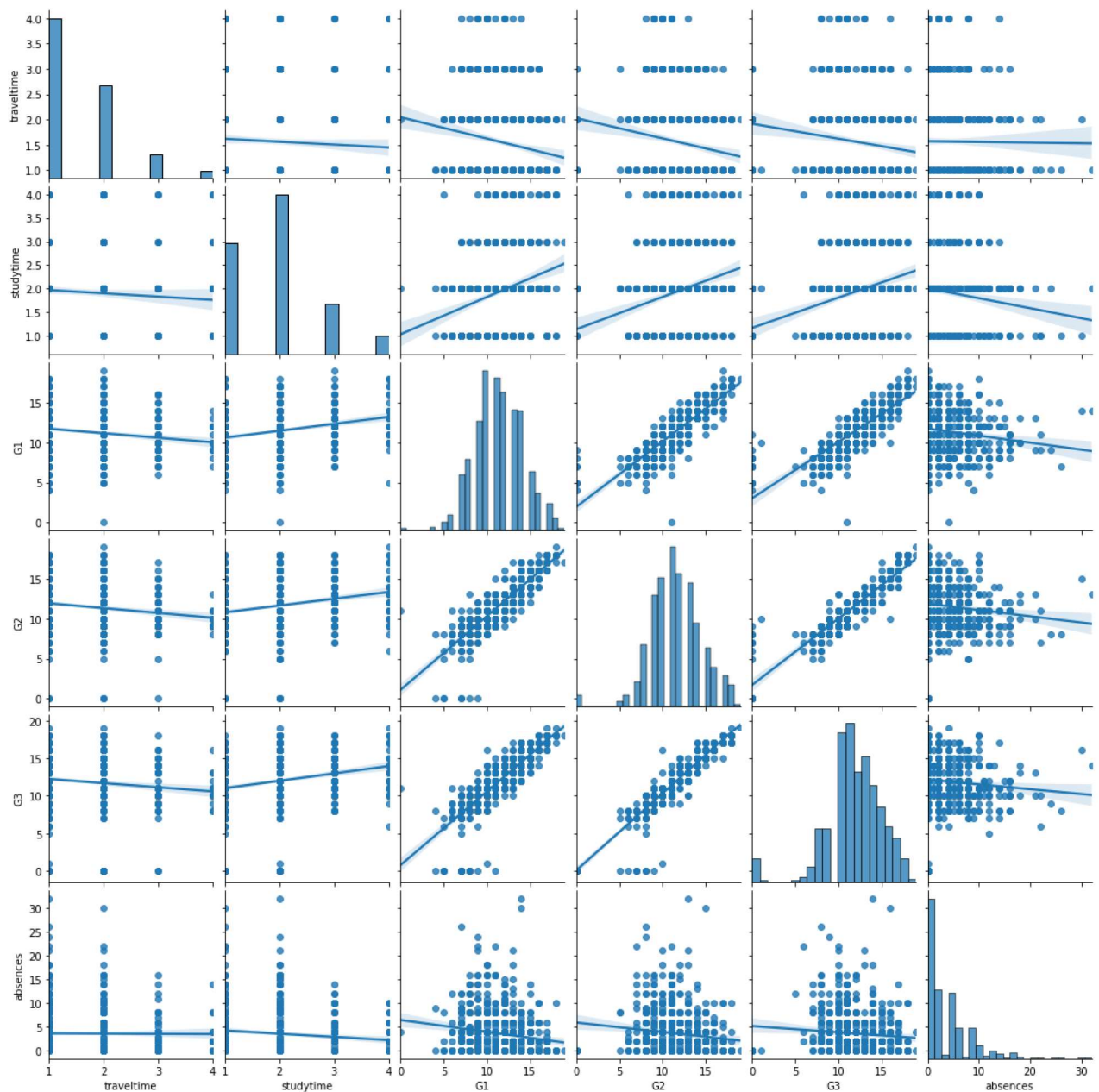
df1a.corr()
```

```
Out[372]:
```

	traveltime	studytime	G1	G2	G3	absences
traveltime	1.000000	-0.063154	-0.154120	-0.154489	-0.127173	-0.008149
studytime	-0.063154	1.000000	0.260875	0.240498	0.249789	-0.118389
G1	-0.154120	0.260875	1.000000	0.864982	0.826387	-0.147149
G2	-0.154489	0.240498	0.864982	1.000000	0.918548	-0.124745
G3	-0.127173	0.249789	0.826387	0.918548	1.000000	-0.091379
absences	-0.008149	-0.118389	-0.147149	-0.124745	-0.091379	1.000000

Again, we see a negative correlation between travel time and student grades at all three measured performance checkpoints. However, when compared to Math performance, we see a notably stronger positive correlation between study time and Portuguese grades. In this case, the magnitudes of the study time and grades correlations are larger than the magnitudes of the corresponding travel time and grade correlations. Here, absences are more strongly negatively correlated with Portuguese course grades, unlike absences and math grades in the previous analysis. However, absences and commute time are not significantly correlated, aside from a very slight negative correlation, which is consistent across both data sets.

```
In [373]: # A pairplot was drawn using seaborn scatterplots
# The kind="reg" parameter was used to add linear regression models to the scatter plots,
# the visualization of any correlations
h = sns.pairplot(df1a, kind="reg")
```



In [374]: `# Save pairplot for use in our project report`

`h.savefig("por_output.png")`

As we can see from the table above, travel time is negatively correlated with study time and course grades at all grading checkpoints.

However, unlike in the Math grade data, travel time does not have a stronger negative correlation with final two course grades than the positive correlation that study time has with these final course grades. For these data sets, it appears that time spent studying for Portuguese had a stronger positive correlation with course grades than time spent studying for Math grades.

We will discuss our findings, and the limitations of this project, further in the associated presentation.

References:

Data set source (linked above in the notebook and which includes the attribute information):

<https://archive.ics.uci.edu/ml/datasets/Student+Performance>