PSTAT 100 Final Project: Analyzing Academic Performance Factors

Yongheng Zan (Ryan), Brendan Morrison, Gavin Tieng, and Andrew Pascual

Author Contributions

Yongheng Zan (Ryan): abstract, background, datasets section

Brendan Morrison: distribution plots, aims, results/methods section

Gavin Tieng: regression analysis, results/methods section

Andrew Pascual: background, aims, discussion

0. Abstract

Student success in school can often lead to success in the real world and vice versa. Being able to predict a students success and even influencing them in a positive way would lead to a general increase in grades and overall wellbeing of students. In order to go about this goal we must first find the factors that have the biggest impact on a students performance in school. Once we are able to distinguish which factors have the greatest influence on student grades, we can predict whether a student will pass or fail the class. After conducting analysis on our dataset, we found that the number of previous class failures and gender were the biggest factors in predicting whether a student will pass or fail a given class. To support these findings, we estimated the differences in means between these variables and overall grade, and found that only previous failures and gender led to a statistically significant difference in whether the student passes or fails a class.

1. Introduction

Our project uses the Student Performance Data Set (https://archive.ics.uci.edu/ml/datasets/Student+Performance) from the UCI Machine Learning Repository to explore the relationships between academic achievement and student demographic information.

1a) Background

The Student Performance dataset that we chose contains information on secondary education students from two different schools in Portugual. It primarily consists of information pertaining to class grades, alongside demographic and personal information for each student. Additionally, our data was collected through school reports and questionnaires, with some variables being coded on specific scales (such as 1-5 for "very bad" to "excellent").

Student achievenment and education is a very important aspect of our society; if students are unable to effectively learn and be succesful in school, they will likely experience many difficulties transitioning to the real world when they grow up. Through analyzing this dataset we can learn which factors make the biggest impact on student grades. Thus, these findings can help us predict student success in the future and create an overall better system or environment for students to achieve at a higher level. Although student achievement can be attributed to past grades, there is also a large chance that other outside variables are having an affect as well. With the analysis of this data we hope to provide a clear picture of what factors are hurting students and which ones are helping students in their academic achievement.

1b) Aims

Our primary aim is to explore what factors have the greatest impact (both positive and negative) on a given student's academic grades. Additionally, we will analyze these factors to see if they have a significant impact on predicting whether a given student will pass or fail a class.

To approach our questions, we will begin by visualizing the variables in our dataset and identifying which ones are most correlated to a student's overall grade. Next, we will plot these variables against student grades and separate plots by different variables (such as gender or overall health), in order to identify any patterns or interesting relationships in the data. Finally, we will use a model to estimate whether these factors actually have a significant impact on whether or not a student passes or fails a class.

For our analysis we chose the variables `age`, `sex`, `health`, and `failures` as predictors for `overall_grade`. After creating exploratory plots, the data pointed to age, sex, and health having a lower impact on difference in overall grade, while number of past failures had a larger impact. This indication was partially supported by the linear model that we constructed. Through the model, we found that `age` and `health` had little impact on whether a student passed or failed a class (`overall_grade`) while `failures` and `sex` had a significant impact.

The main questions we aimed to answer in this project are as follows:

1. What features have the greatest impact on student grades?

2. Does the data suggest that these features have a significant impact on predicting whether a student will pass or fail?

2. Materials and methods

2a) Datasets

The Student Performance Dataset contains both student achievement and demographic information on secondary education students from Portugual. The data itself was collected from school reports and surveys given to students from a math class and a portguese class within two different schools.

The table below contains information on the variables of interest for our analysis:

Name	Variable description	Type	Units of measurement
sex	student's sex ('F' - female or 'M' - male)	str	gender
age	student's age (from 15 to 22)	numeric	age
failures	number of past class failures (n if 1<=n<3, else 4)	numeric	number of classes
health	current health status (from 1 - very bad to 5 - very good)	numeric	1 to 5
por_G1	first period Por grade (from 0 to 20)	numeric	academic score
por_G2	second period Por grade (from 0 to 20)	numeric	academic score
por_G3	final Por grade (from 0 to 20, output target)	numeric	academic score
mat_G1	first period math grade (from 0 to 20)	numeric	academic score
mat_G2	second period math grade (from 0 to 20)	numeric	academic score
mat_G3	final math grade (from 0 to 20, output target)	numeric	academic score

The format of the data is as follows:

In [1]:
```
import pandas as pd
data = pd.read_csv('tidy_data.csv')
data.head()
```

Out[1]:

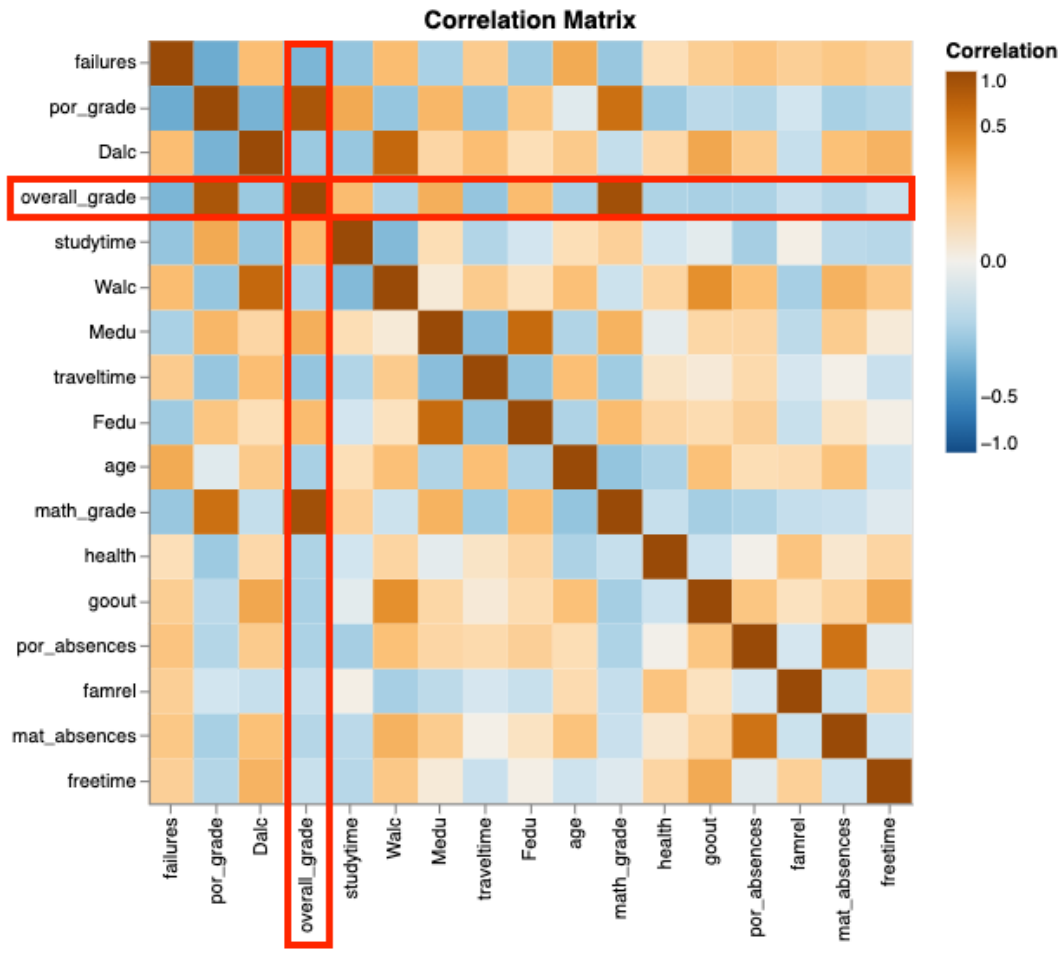
	sex	age	failures	health	por_G1	por_G2	por_G3	mat_G1	mat_G2	mat_G3
0	F	18	0	3	0	11	11	5	6	6
1	F	17	0	3	9	11	11	5	5	6
2	F	17	0	3	9	11	11	8	8	9
3	F	15	0	5	14	14	14	15	14	15
4	F	16	0	5	11	13	13	6	10	10

2b) Methods

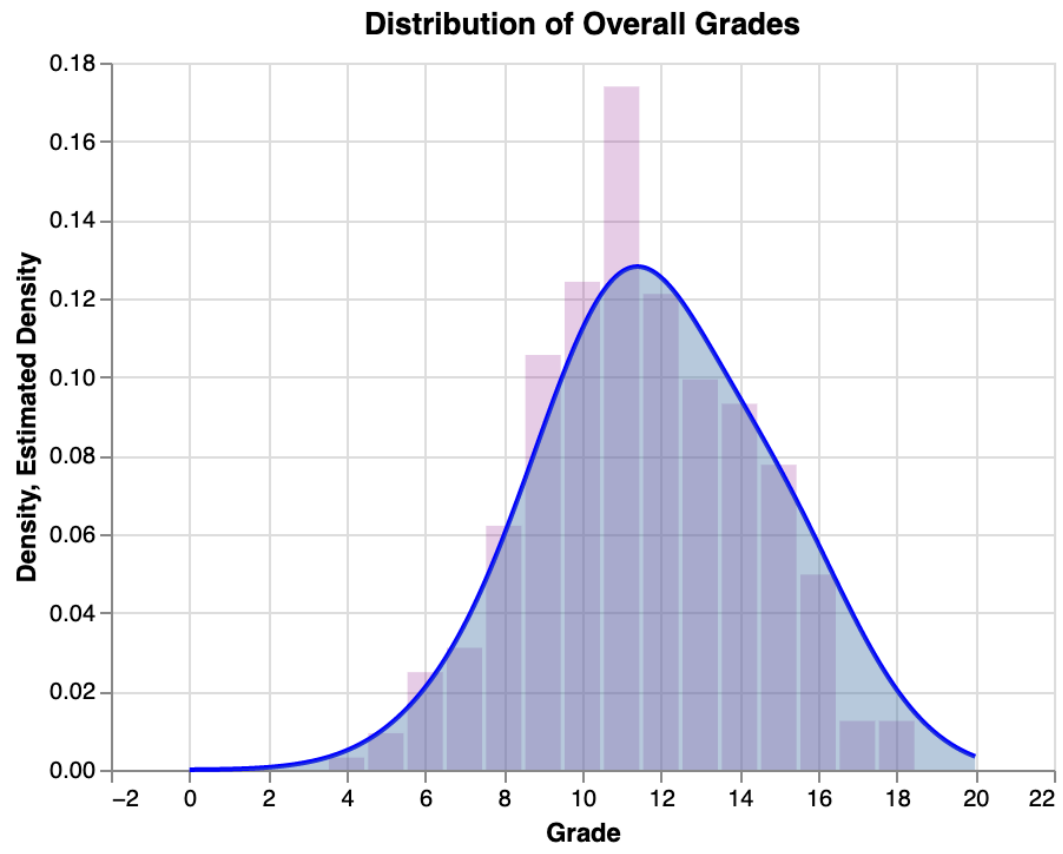
We began our research by conducting exploratory analysis, which involved creating a heatmap to get an idea of what variables were most correlated with overall academic grade. Once we identified a few variables of interest (`age` , `sex` , `health` , `failures`), we visualized the distribution of the overall grades grouped by these variables, using histograms and stacked density curves. Next, we encoded the `overall_grade` variable into a new feature that represented whether a student passed or failed, and used the following multiple linear regression model to estimate each coefficient and analyze the prediction capability of these features on student performance:

$$(\text{Pass/Fail}_i) = \beta_0 + \beta_1(0\text{ Failures})_i + \beta_2(1\text{ Failures})_i + \cdots + \beta_4(3\text{ Failures})_i + \beta_5(\text{Male})_i + \beta_6(\text{Very Bad Health})_i + \cdots + \beta_{10}(\text{Very Good Health})_i + \beta_{11}(\text{Age Under 18})_i + \epsilon_i$$

3. Results

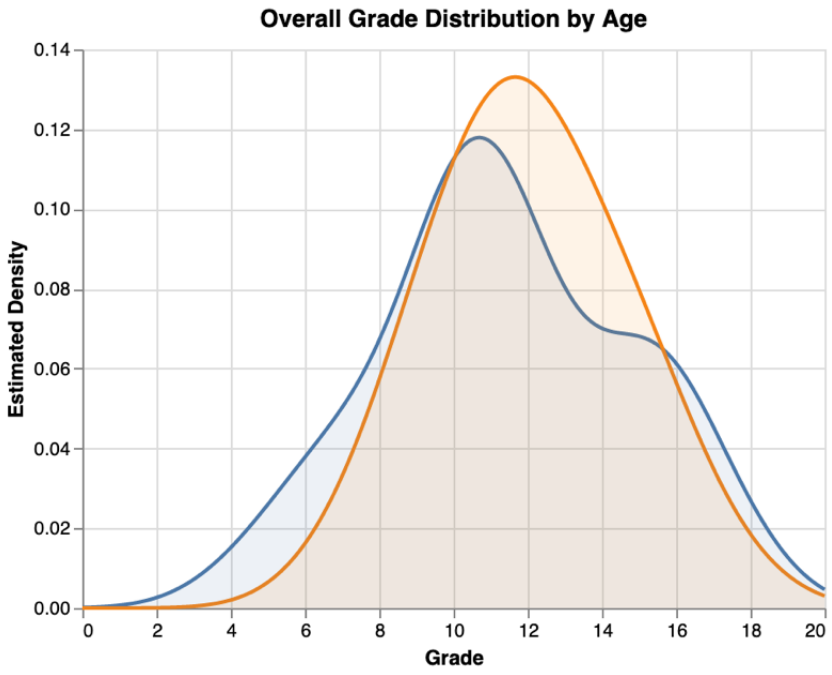


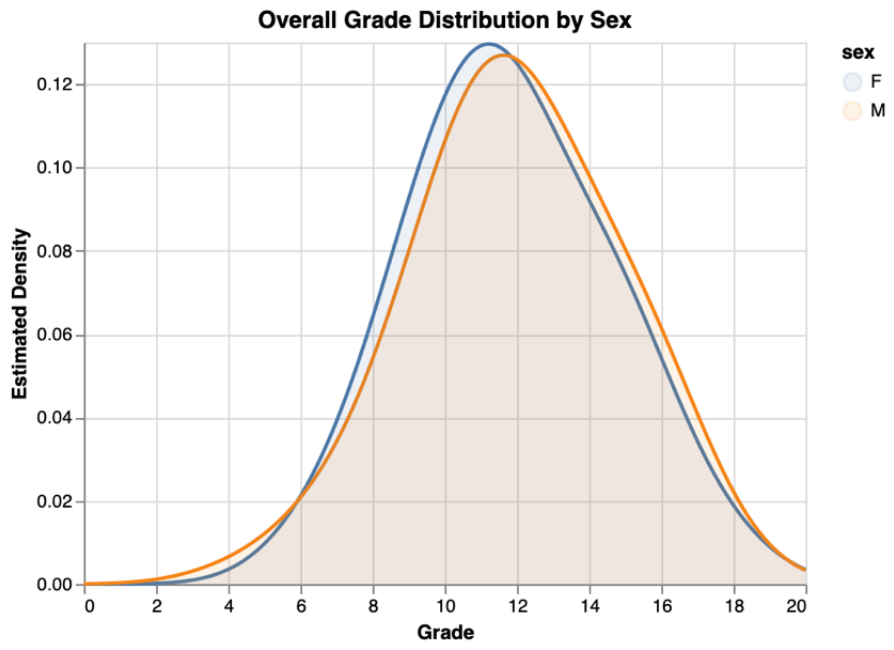
Using the above heatmap, we identified some variables that could potentially have an impact on the `overall_grade` variable. Namely, `failures` and `health`. In addition, we decided to include `age` and `sex`, as they would allow us to create visualizations that incorporated student demographic information. We did not include the `por_grade` and `math_grade` variables, as they were used to create the `overall_grade` variable. Next, we examined the estimated density of `overall_grade`:



Based on the visualization above, student grades appear to follow a normal distribution. A large number of grades fall between 10 and 12- which is right on the cutoff between a "pass" and a "fail". The overall grade distribution shows a median grade of about 11/20. The bell curve shape, being relatively evenly distributed above 10 and below 10 indicates a roughly equal amount of observations with a passing overall grade and a failing overall grade. The even distribution between pass and fail will help in the accuracy of our model's prediction accuracy.

Next, we created separate density curve estimate plots for the `overall_grade` variable, with each plot being sorted by one of our variables of interest. Below are the `overall_grade` plots, sorted by our demographic variables, `age` and `sex`:



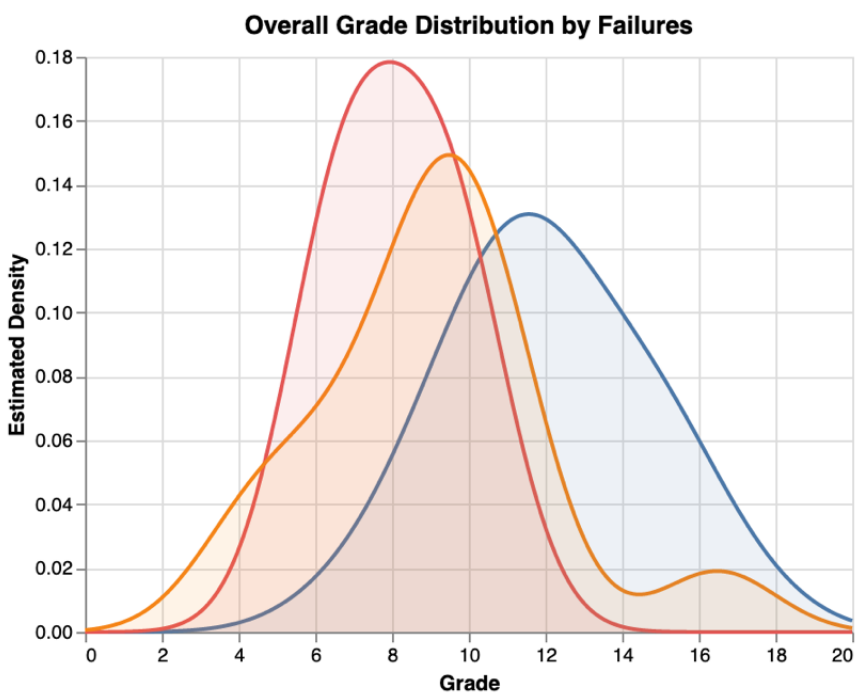


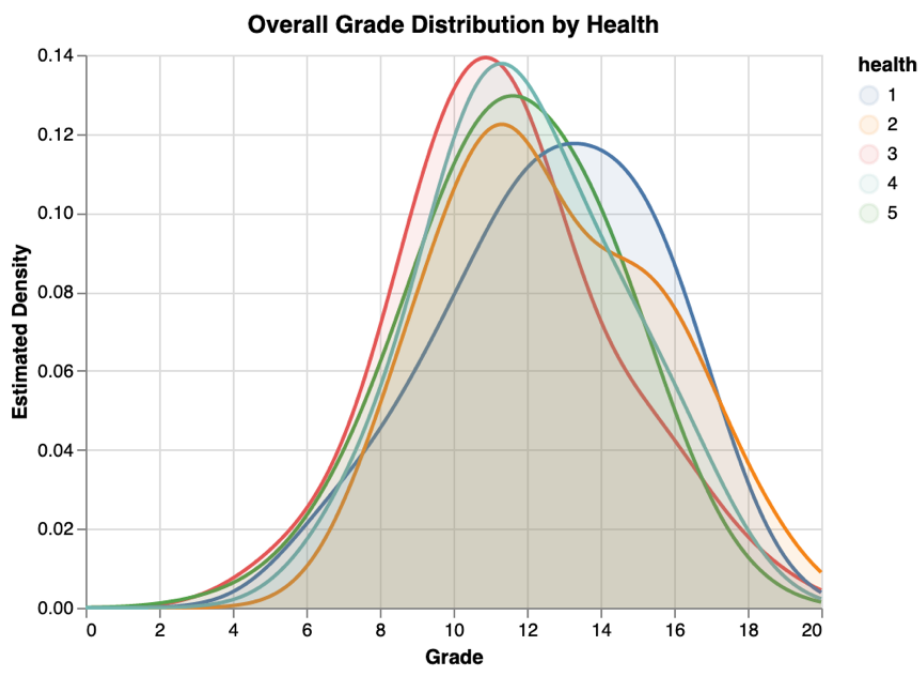
For the figure on the left ("Overall Grade Distribution by Age"), we can see that a slightly larger proportion of students fall into the over 18 cohort. In addition, the students in the over 18 cohort have a higher mean grade compared to the students in the under 18 cohort. There were approximately 251 students in the under 18 cohort, and 71 students in the over 18 cohort, so there is reason to believe that the under 18 cohort is a slightly better representation of the population.

The figure on the right ("Overall Grade Distribution by Sex") shows us that a student's gender doesn't seem to affect their grades. While males have a slightly higher average grade, it is very similar to that of females. This was particularly interesting, as our later analysis indicated that differences in gender have a significant impact on differences in overall grade.

Our overall grade distribution plots, when grouped by different variables, can show how the overall grade is affected by those variables. We can see in these plots that students under 18 and students that are female have a slightly lower median overall grade. Similar to the distribution of the overall grade, the bell shaped curves of these distributions will help our model give more accurate predictions, as the data does not seem to be skewed in any way.

Similarly, we created 2 more density curve estimate plots for the `overall_grade` variable, but this time, sorted them by the variables that we selected after examining the heatmap (`failures` and `health`):





The figure on the left ("Overall Grade Distribution by Failures") shows us that students who failed multiple classes had, on average, lower grades. This trend was consistent for those who had 1 previous failure and for those who had 0 previous failures- the highest grades typically fell in the 0 failures category. This is logical- it can be inferred that a student who has never failed a class will likely be getting higher grades, on average.

The figure on the right ("Overall Grade Distribution by Health") has the highest median overall grade for students with 1 (very bad) health and similar distributions for the other health categories. The overall grade distribution, when grouped by number of failures, seems to have the greatest difference between the levels. A low number of failures in previous classes seems to correlate with a higher overall grade, and a high number of failures in previous classes seems to correlate with a lower overall grade.

Finally, we used a multiple linear regression model to help estimate how `age`, `sex`, `health`, and `failures` affect `overall_grade`. The parameters in our model represent the differences in means between student age (when gender, health, and failures are fixed), student gender (when age, health, and failures are fixed), student health (when age, gender, and failures are fixed), and the number of student failures (when age, gender, and health are fixed).

We transformed the `overall_grade` variable into one called `grade_pass_fail`, which was 1 if a student passed the class and 0 if they did not. The cutoff for a pass versus a fail was based on a 0 to 20 scoring scale- if a student scored between 0 and 10, they failed the class- if they scored between 11 and 20, they passed the class. We found that approximately 246 students passed and 76 students failed.

After encoding all of our variables as indicators using one-hot encoding, we added an intercept, defined a response variable, and fit the model. Below is the parameter estimation table, with columns that contain information on each predictor's standard error, plus/minus 2 standard deviations, and whether any differences between predictor levels are statistically significant:

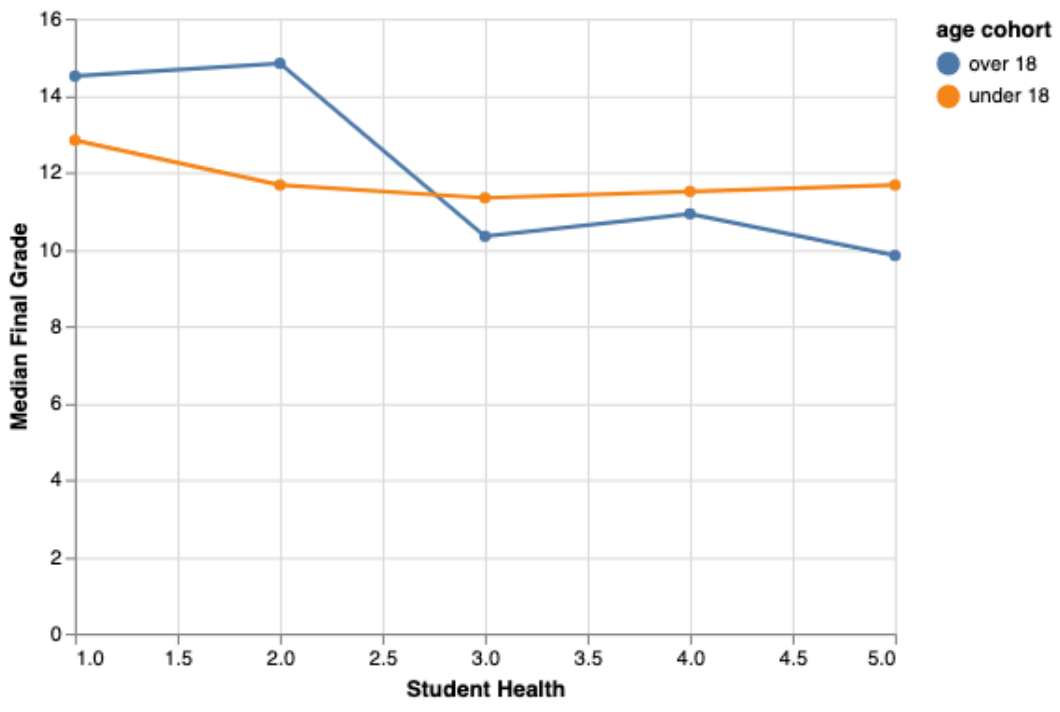
```
Mean Squared Error = 0.16164
```

	Group	Coefficient Estimates	Standard Error	-2*SE	+2*SE	Significant
0	failures	0.744125	0.076418	0.591288	0.896961	Yes
1	sex_M	-0.301551	0.059469	-0.420489	-0.182613	Yes
2	health_2	0.056802	0.046438	-0.036075	0.149679	No
3	health_3	-0.035577	0.093093	-0.221762	0.150608	No
4	health_4	-0.109714	0.078037	-0.265789	0.046361	No
5	health_5	-0.006026	0.082359	-0.170744	0.158692	No
6	age cohort_under 18	-0.099716	0.072831	-0.245378	0.045946	No
7	intercept	0.107754	0.055633	-0.003513	0.219020	No
8	error variance	0.165755	NaN	NaN	NaN	NaN

We found that differences in the `failures` variable were significant, which indicates that whether a student passes or fails a class is influenced by how many previous failures they had (regardless of which class).

Another surprising finding was that differences in the `sex` variable were significant- it appears that a student's gender affects whether a student passes or fails a class. While the underlying cause of this result is more difficult to explain, it is plausible that it is due to sampling bias that occurred when the data was collected.

Lastly, we found that none of the differences in the `age` and `health` variables were significant, indicating that a student's age and their health status do not have a large impact on whether they pass or fail the class. This can be further explained by the following visualization:



We can see that the median final grade for students who are under the age of 18 is roughly the same for all levels of health, falling approximately in the 11-13 range, which is right on the line between passing or failing the class. The median final grade for students over the age of 18 has a slightly wider range, falling between 10 and 15, which is also on the line between passing and failing. However, there does not appear to be a clear relationship between health and overall grade- while better health seems like it would correlate to higher grades, this plot actually showed us the opposite.

4. Discussion ¶

There are numerous variables that can affect a students academic performance. However, in analyzing our dataset we have found that the most positive correlation comes mainly from previous good grades. With slightly less correlation are the study time, and parents level of education. Good study habits are seemingly passed down from a students parents which leads to higher grades. Contrarily, the factor with the most negative correlation on overall grades was travel time from home to school. Travel time could have such an effect as it forces the students to physically have less time for school work or studying. Additionally, alcohol consumption was the next leading negative correlation to student grades. However, this comes as no surprise, alcohol consumption is well known for having an inverse relationship with student success.

Overall, these several variables of a students life and choices do seem to be a good indicator of whether they will pass or fail in school. Thus, Students with more negative factors in their life tend to have more failures, and students with more positive factors tend to have no failures. However, these factors do not always determine a students success, there are some outlier students who pass all their classes while dealing with more negative variables, likewise there are students who still struggle when they displayed more positive variables. Furthermore, if time had allowed, it would have been beneficial to be able to delve deeper into our dataset and further our findings. Such as, adding other factors to deepen our analysis, or looking at a different main objective other than just grades or pass/fail. In conclusion, our research and analysis provides a good baseline to determine academic success of students, but it can always be improved or changed to meet different criteria.

In [ ]: