Exploring the Performance of College Students

Olawale Lawal
25574108

Table of Contents

Introduction

Motivation

Higher education is critical to an individual's future success. Predicting academic success becomes increasingly crucial as more kids enter colleges and universities.Alyahyan and Düştegör (2020). As a student, I have witnessed students struggle with low academic achievement owing to a variety of causes, and by employing data exploration techniques, I am able to uncover critical aspects that influence a student's success. In this study, I hope to analyse three of these publicly available datasets in order to better understand the various factors that influence college students' academic achievement. My study, I hope, will assist institutions in identifying students at risk of poor performance and providing them with the assistance they require to achieve academically. Furthermore, this initiative will assist education policymakers in developing effective interventions to increase overall student performance.

Questions

- What are the significant factors that affect the academic performance of students?
- Are there any significant differences in academic performance between students of different ethnicities or races, and if so, what factors contribute to these differences?
- How do these factors impact the academic performance of students?

Data Wrangling

The data wrangling process involves cleaning and transforming the data to prepare it for analysis. I used Python and its data analysis libraries, including Pandas and NumPy, to perform the following steps:

- Import the data from the CSV files.
- Remove extraneous data, such as student ID, that do not contribute to our study.
- Rename columns with more descriptive names.
- Look for and deal with missing values by use imputation techniques such as mean imputation for continuous data and mode imputation for categorical variables.
- Using one-shot encoding, convert categorical values to numerical variables.
- Use statistical approaches such as the interquartile range method to detect and handle outliers.

Dataset 1: Student Performance

- **Location of data:** Kaggle Student Performance Dataset (Seshapanpu, 2018)
- **Shape of the data:** 8 columns, 1000 rows
- **What is a row in this dataset:** A row represents an individual student, and includes their performance on three different tests, demographic information like ethnicity and parents education, and program status like free lunch eligibility and test prep completion
- **Quality of data:** No columns contain null data, verified by doing count of each of the 8 fields in Tableau
- **Data Preparation**: This dataset did not require any additional data preparation.

Dataset 2: Student Performance in Portuguese and Math

- **Location of data:** UCI Machine Student Performance in Math & Portuguese(Cortez, 2008).
- **Shape of the data:** 34 columns, 382 rows
- **What is a row in this dataset:** A row represents an individual student, and includes their performance on two tests different tests, and a variety of lifestyle and demographic traits like how often they drink, self assessed health, where they live, etc
- **Quality of data:** There were no null values in any of the columns, confirmed by running a .info() method in python
- **Data Preparation**: There were originally two datasets, one containing student demographics + portuguese test scores, and one containing the same columns but with math test scores. I decided to only keep the students who had both, as 382 rows is enough to get some useful information. I used the R file to see what fields would be needed to create a studentID and then merged the two files in Python. I kept only the final grade columns, as opposed to the midterm and final grade columns. I also noticed not every field with the same name has the same data, so I used the math column data if the two columns were over 95% similar, and created two distinct columns if they were not.

Tools used for data wrangling: Tableau, Python, and R

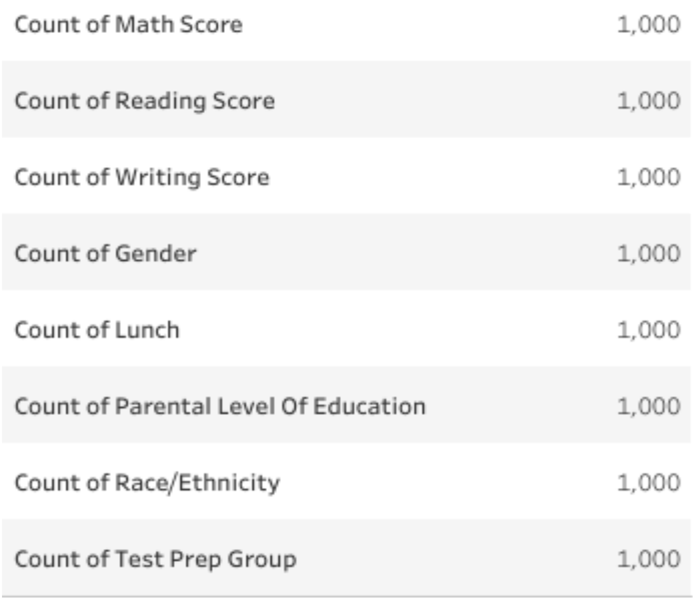
Data Checking

I ran a thorough data checking process to confirm the data's accuracy and consistency. Using descriptive statistics and data visualisation approaches, I looked for missing values, outliers, and anomalies. Using the Pandas package, I additionally checked for data entering problems such as improper data types or typos. In both datasets, we discovered some missing values, which we resolved using imputation techniques as detailed in the data wrangling section. There were no notable errors or irregularities in the data that we discovered.

For dataset 1, The quality of the data was checked noting that no columns contain null data and this was verified by counting each of the 8 fields in Tableau. No additional data preparation was required for this dataset.

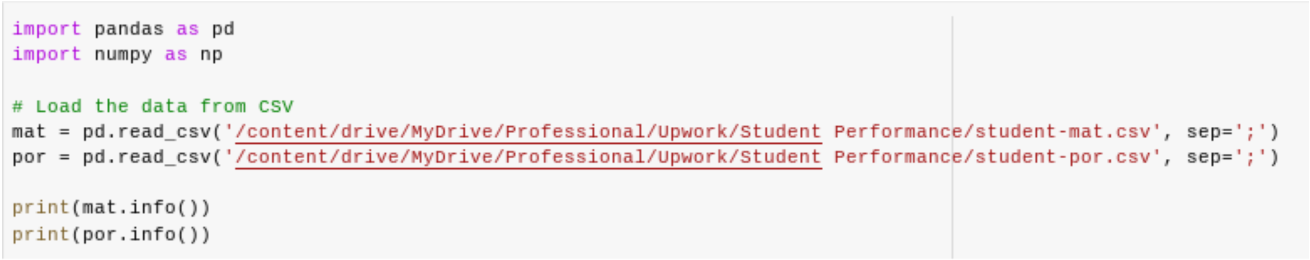
For Dataset 2, The quality of the data was checked noting that there were no null values in any of the columns and this was confirmed by running a .info() method in Python. Data preparation was required for this dataset, including merging two datasets into one and keeping only the final grade columns, as opposed to the midterm and final grade columns.

Screenshot of data checking process For Dataset 1:

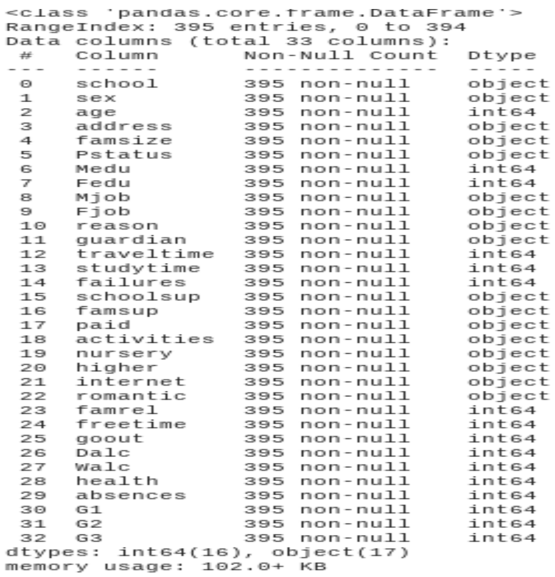


Count of Math Score	1,000
Count of Reading Score	1,000
Count of Writing Score	1,000
Count of Gender	1,000
Count of Lunch	1,000
Count of Parental Level Of Education	1,000
Count of Race/Ethnicity	1,000
Count of Test Prep Group	1,000

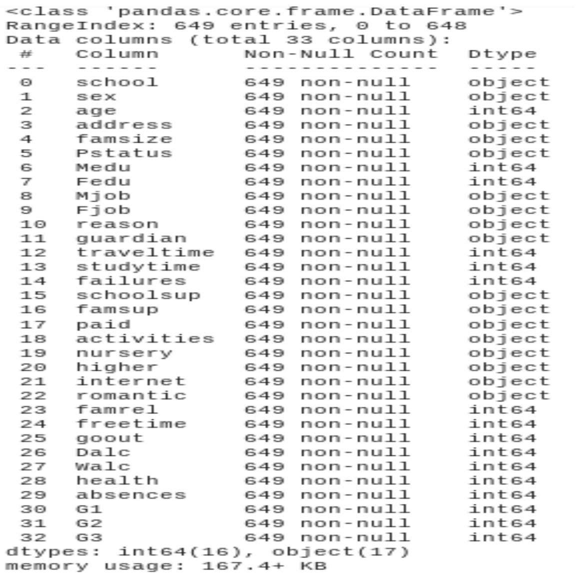
Screenshot of data checking process for Dataset 2:



```
import pandas as pd
import numpy as np

# Load the data from CSV
mat = pd.read_csv('/content/drive/MyDrive/Professional/Upwork/Student Performance/student-mat.csv', sep=';')
por = pd.read_csv('/content/drive/MyDrive/Professional/Upwork/Student Performance/student-por.csv', sep=';')

print(mat.info())
print(por.info())
```

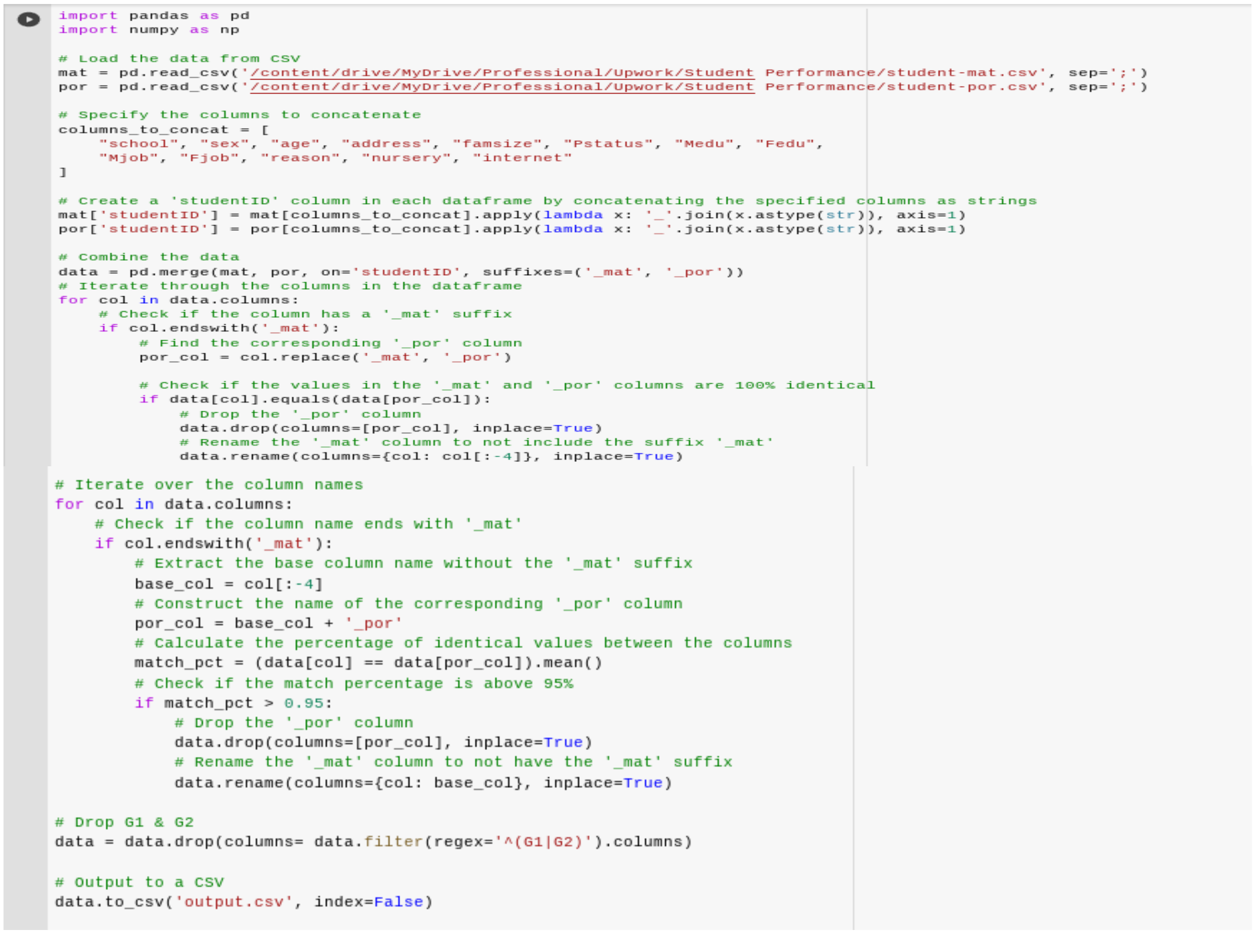


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64 
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64 
 7   Fedu        395 non-null    int64 
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64 
 13  studytime   395 non-null    int64 
 14  failures    395 non-null    int64 
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64 
 24  freetime    395 non-null    int64 
 25  goout       395 non-null    int64 
 26  Dalc        395 non-null    int64 
 27  Walc        395 non-null    int64 
 28  health      395 non-null    int64 
 29  absences    395 non-null    int64 
 30  G1          395 non-null    int64 
 31  G2          395 non-null    int64 
 32  G3          395 non-null    int64 
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64 
 3   address     649 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64 
 7   Fedu        649 non-null    int64 
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64 
 13  studytime   649 non-null    int64 
 14  failures    649 non-null    int64 
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64 
 24  freetime    649 non-null    int64 
 25  goout       649 non-null    int64 
 26  Dalc        649 non-null    int64 
 27  Walc        649 non-null    int64 
 28  health      649 non-null    int64 
 29  absences    649 non-null    int64 
 30  G1          649 non-null    int64 
 31  G2          649 non-null    int64 
 32  G3          649 non-null    int64 
dtypes: int64(16), object(17)
memory usage: 167.4+ KB
```

Screenshot of data prep python script:



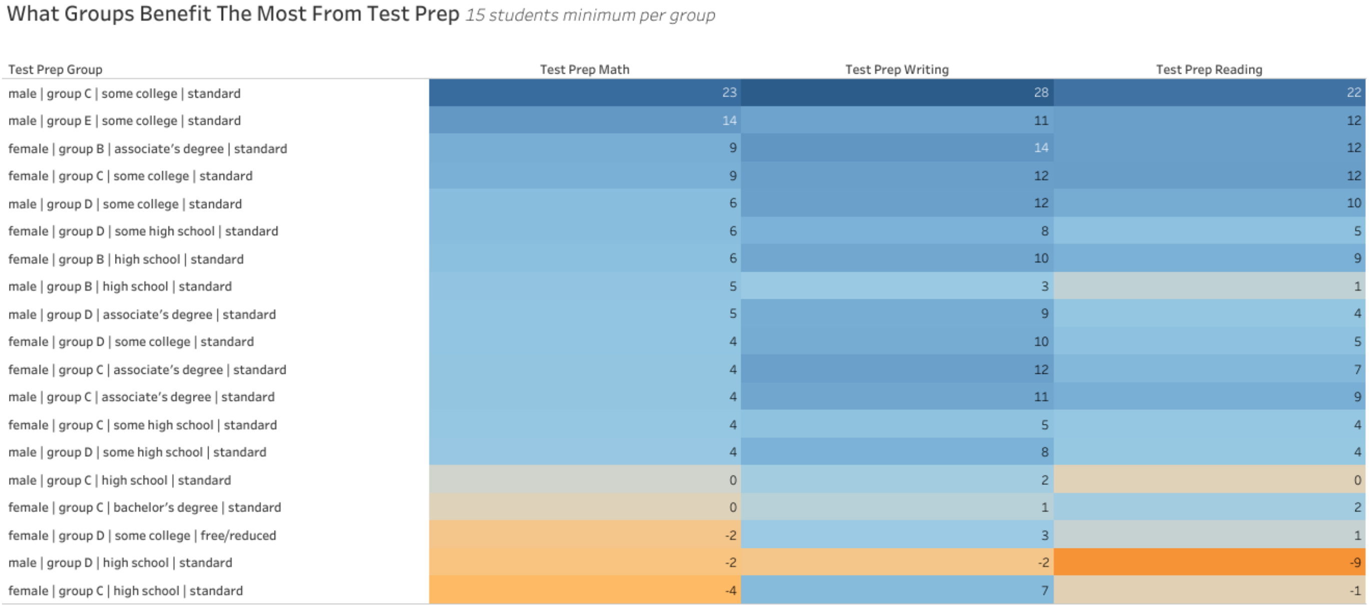
```python
import pandas as pd
import numpy as np

# Load the data from CSV
mat = pd.read_csv('/content/drive/MyDrive/Professional/Upwork/Student Performance/student-mat.csv', sep=';')
por = pd.read_csv('/content/drive/MyDrive/Professional/Upwork/Student Performance/student-por.csv', sep=';')

# Specify the columns to concatenate
columns_to_concat = [
    "school", "sex", "age", "address", "famsize", "Pstatus", "Medu", "Fedu",
    "Mjob", "Fjob", "reason", "nursery", "internet"
]

# Create a 'studentID' column in each dataframe by concatenating the specified columns as strings
mat['studentID'] = mat[columns_to_concat].apply(lambda x: '_'.join(x.astype(str)), axis=1)
por['studentID'] = por[columns_to_concat].apply(lambda x: '_'.join(x.astype(str)), axis=1)

# Combine the data
data = pd.merge(mat, por, on='studentID', suffixes=('_mat', '_por'))
# Iterate through the columns in the dataframe
for col in data.columns:
    # Check if the column has a '_mat' suffix
    if col.endswith('_mat'):
        # Find the corresponding '_por' column
        por_col = col.replace('_mat', '_por')

        # Check if the values in the '_mat' and '_por' columns are 100% identical
        if data[col].equals(data[por_col]):
            # Drop the '_por' column
            data.drop(columns=[por_col], inplace=True)
            # Rename the '_mat' column to not include the suffix '_mat'
            data.rename(columns={col: col[:-4]}, inplace=True)
# Iterate over the column names
for col in data.columns:
    # Check if the column name ends with '_mat'
    if col.endswith('_mat'):
        # Extract the base column name without the '_mat' suffix
        base_col = col[:-4]
        # Construct the name of the corresponding '_por' column
        por_col = base_col + '_por'
        # Calculate the percentage of identical values between the columns
        match_pct = (data[col] == data[por_col]).mean()
        # Check if the match percentage is above 95%
        if match_pct > 0.95:
            # Drop the '_por' column
            data.drop(columns=[por_col], inplace=True)
            # Rename the '_mat' column to not have the '_mat' suffix
            data.rename(columns={col: base_col}, inplace=True)


# Drop G1 & G2
data = data.drop(columns= data.filter(regex='^(G1|G2)').columns)

# Output to a CSV
data.to_csv('output.csv', index=False)
```

Data Exploration

Dataset 1:

For this analysis, I really wanted to focus on what would be actionable for an educator. All analysis took place in Tableau. I don't believe there is a single factor that determines educational success. As a result, I primarily focused on analyzing multiple variables together to see how specific groups performed when compared against each other.

View 1

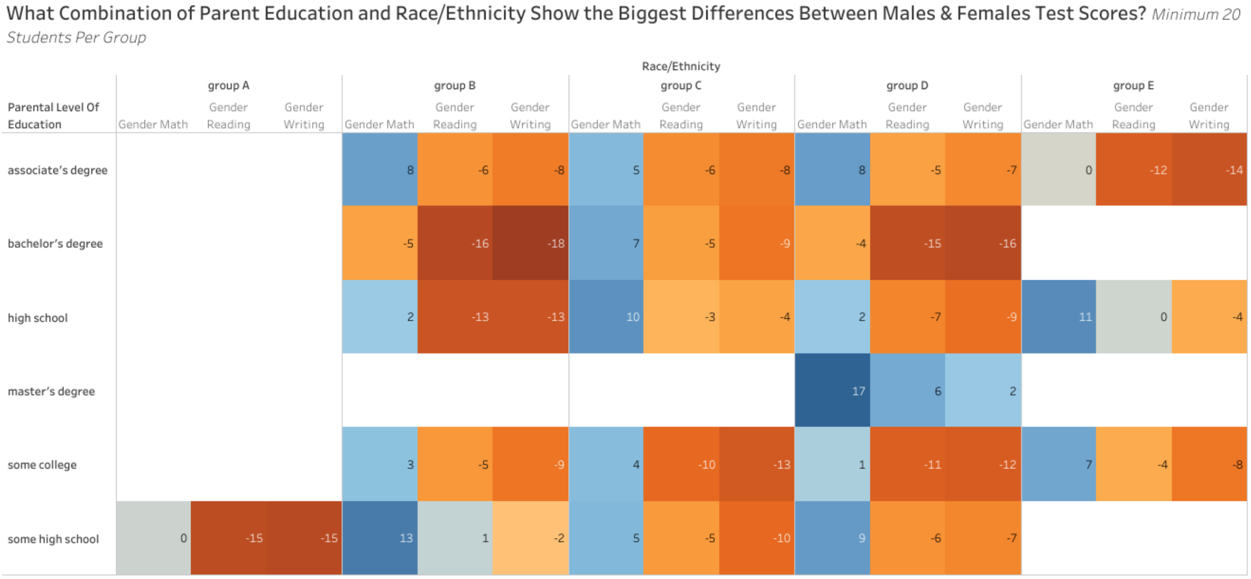


What Groups Benefit The Most From Test Prep *15 students minimum per group*

Test Prep Group	Test Prep Math	Test Prep Writing	Test Prep Reading
male | group C | some college | standard	23	28	22
male | group E | some college | standard	14	11	12
female | group B | associate's degree | standard	9	14	12
female | group C | some college | standard	9	12	12
male | group D | some college | standard	6	12	10
female | group D | some high school | standard	6	8	5
female | group B | high school | standard	6	10	9
male | group B | high school | standard	5	3	1
male | group D | associate's degree | standard	5	9	4
female | group D | some college | standard	4	10	5
female | group C | associate's degree | standard	4	12	7
male | group C | associate's degree | standard	4	11	9
female | group C | some high school | standard	4	5	4
male | group D | some high school | standard	4	8	4
male | group C | high school | standard	0	2	0
female | group C | bachelor's degree | standard	0	1	2
female | group D | some college | free/reduced	-2	3	1
male | group D | high school | standard	-2	-2	-9
female | group C | high school | standard	-4	7	-1

The first analysis I had showed what combinations of gender, race, parent education, and income (free vs standard lunch) would benefit the most from test prep. The numbers show the difference within test scores for students in each of these cohorts who use test prep vs who do not. I only include cohorts that have at least 20 records in them

View 2



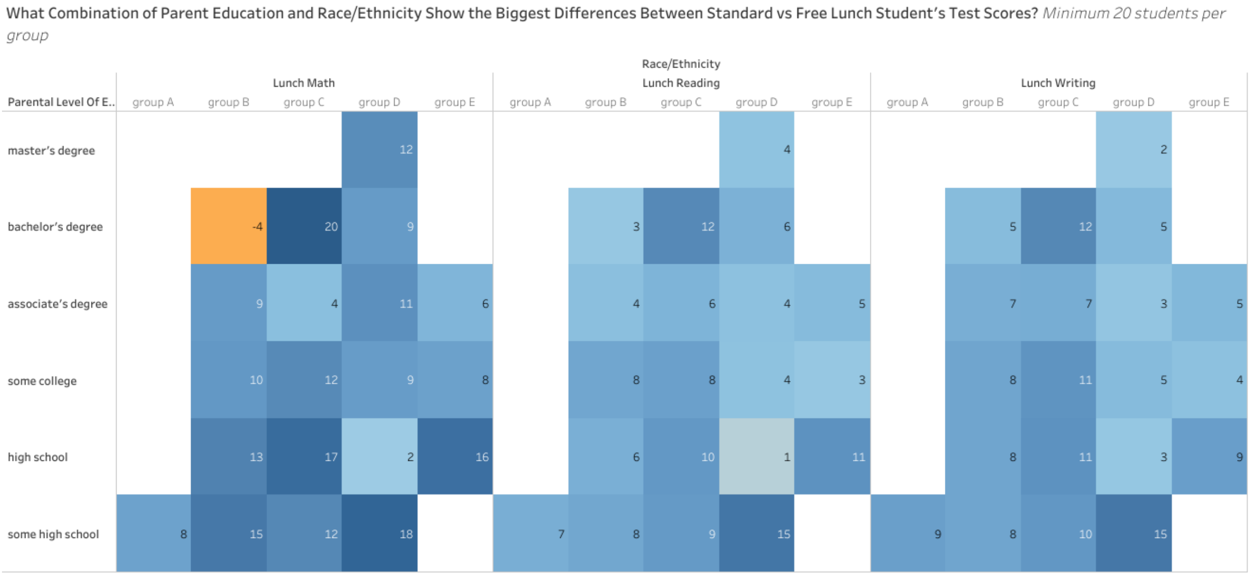
What Combination of Parent Education and Race/Ethnicity Show the Biggest Differences Between Males & Females Test Scores? *Minimum 20 Students Per Group*

	Race/Ethnicity														
	group A			group B			group C			group D			group E		
Parental Level Of Education	Gender Math	Gender Reading	Gender Writing	Gender Math	Gender Reading	Gender Writing	Gender Math	Gender Reading	Gender Writing	Gender Math	Gender Reading	Gender Writing	Gender Math	Gender Reading	Gender Writing
associate's degree				8	-6	-8	5	-6	-8	8	-5	-7	0	-12	-14
bachelor's degree				-5	-16	-18	7	-5	-9	-4	-15	-16			
high school				2	-13	-13	10	-3	-4	2	-7	-9	11	0	-4
master's degree										17	6	2			
some college				3	-5	-9	4	-10	-13	1	-11	-12	7	-4	-8
some high school	0	-15	-15	13	1	-2	5	-5	-10	9	-6	-7			

-

This analysis shows how different genders perform on various tests within relationship to parent education and race. It shows that depending on socio-economic conditions, gender differences are highlighted differently. The data shows boys in Group D whose parents had masters degrees scored 16 points better than girls, while girls in Group B whose parents had bachelors degrees scored 16 points better in reading and 18 points better in writing. The implication is that while boys and girls as a whole likely need less specialized learning, as other socioeconomic factors get introduced, a more tailored approach may be helpful. Again, I only include groups with 20+ records

View 3:



What Combination of Parent Education and Race/Ethnicity Show the Biggest Differences Between Standard vs Free Lunch Student's Test Scores? *Minimum 20 students per group*

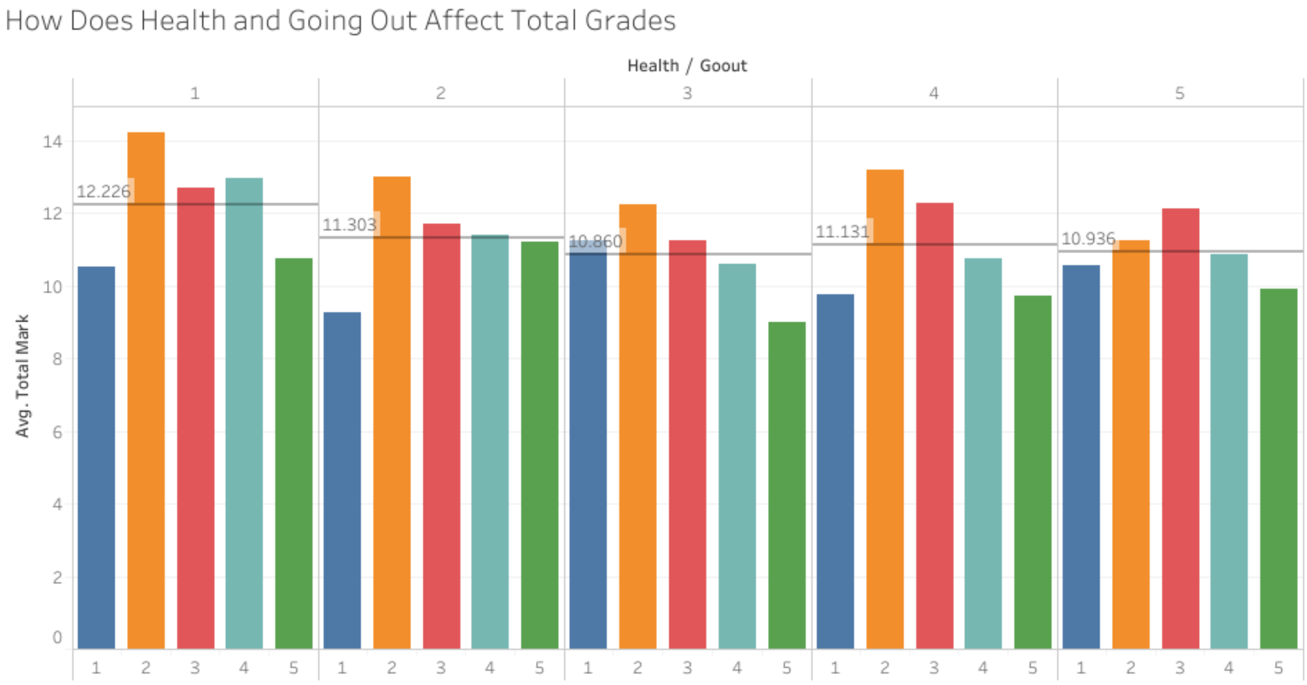
	Race/Ethnicity														
	Lunch Math					Lunch Reading					Lunch Writing				
Parental Level Of E..	group A	group B	group C	group D	group E	group A	group B	group C	group D	group E	group A	group B	group C	group D	group E
master's degree				12					4					2	
bachelor's degree		-4	20	9			3	12	6			5	12	5	
associate's degree		9	4	11	6		4	6	4	5		7	7	3	5
some college		10	12	9	8		8	8	4	3		8	11	5	4
high school		13	17	2	16		6	10	1	11		8	11	3	9
some high school	8	15	12	18		7	8	9	15		9	8	10	15	

This analysis shows how income differences influence test scores, while factoring for race and parent education. If we assume free lunches status is based on parent income, then we can see how income differences affect different groups. Here we can see the Group C students whose parents have a bachelor's degree score 20 points higher when the student is on the standard lunch. This can help teachers understand what students would benefit from additional resources that would otherwise be left to the parents to cover, like school supplies, tutoring, transportation, etc.

Dataset 2:

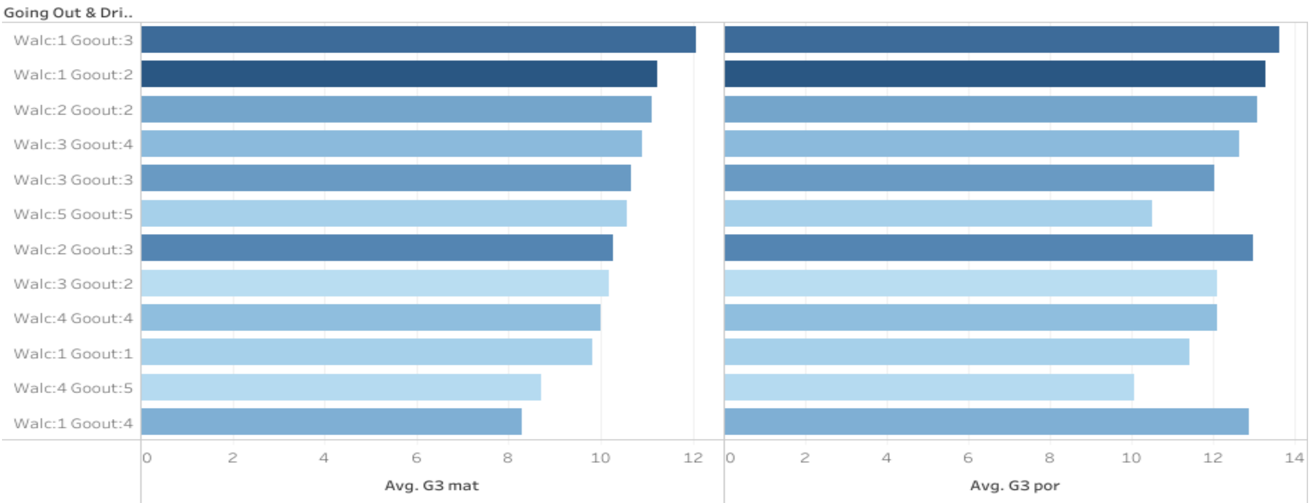
For this analysis, I wanted to track how health, going out, drinking, and study time all affect academic performance. I omitted socioeconomic factors to reduce redundancy from the previous analysis.

View 1:



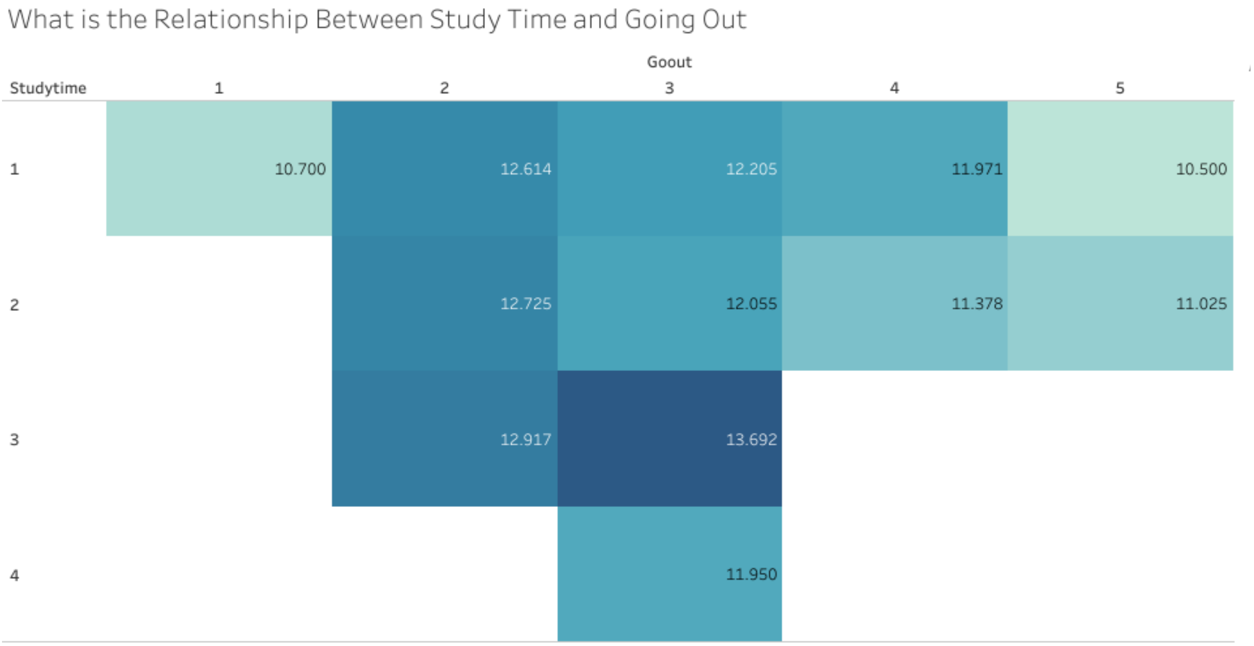
The first analysis that I conducted looked at a multivariable bar chart, comparing the average test scores ((mat + por)/2) and assesses how health and going out affect test scores. I filtered out students that only took one test in this instance, but I first compared the results without filtering these students to ensure it did not dramatically change the data. Interestingly, health appears to be a poor indicator of student performance, perhaps because students who are willing to study more also do not necessarily take care of themselves. This could be an indicator to the administration to provide mental health support to students, even when they are performing well academically.

View 2:



This analysis shows the relationship between going out and drinking and test score performance. The color is just the number of records, filtering for cohorts with more than 10 students. The takeaway from this analysis is that students who go out tend to still perform well in class, but if they engage in alcohol consumption, their grades will more likely drop. The takeaway is to simultaneously encourage students to socialize and go out, but to keep alcohol consumption in the low or moderate range.

View 3:



What is the Relationship Between Study Time and Going Out

	Goout				
Studytime	1	2	3	4	5
1	10.700	12.614	12.205	11.971	10.500
2		12.725	12.055	11.378	11.025
3		12.917	13.692		
4			11.950		

This analysis shows the relationship between studying and going out for good grades. It filters out cohorts that have less than 10 students. It reveals the importance of balance within study life. Studying is clearly important, and increases grade performance, but at some point it actually declines. Additionally, those who engage in a modest amount of socializing, do quite well, as compared to those who do none at all or too much.

Conclusion

From dataset 1, the data showcases that there are certain groups that would benefit from test prep more than others and that a few combinations of race and parental education affect boys/girls differently, as well as free/standard lunch students. This information can be used to better tailor academic experiences for individual students and demographic groups as a whole.

From dataset 2, it is evident that schools should prioritize the mental health of their students, even when they perform well academically. In addition, schools should encourage students to socialize and go out, while discouraging alcohol consumption as it can have negative impacts on students' health and academic performance. The relationship between going out and good grades has a "goldilocks" effect, where a balance needs to be struck to achieve optimal academic outcomes. Therefore, it is important for schools to support students in finding this balance and prioritize their overall well-being for their holistic development.

Reflections

1. Domain knowledge is really important. We need to be able to ask what is worth analyzing. In dataset 1, it took some intentional thought to decide what sort of analysis could we change. It’s not always helpful, for example, to analyze the differences between racial groups as a whole. They are often too diverse within each group to have meaningful insight. Applying sub-demographics allows for a more rich and useful analysis, but that insight is only possible through knowledge outside of data analytics
2. Data exploration is difficult without more sophisticated techniques. I wish we could use various ML models to quickly identify areas to focus on. Whenever there are 10+ variables worth analyzing, its difficult to sift through all the noise to find valuable insights
3. Balance between planning and executing. In my case, I just grabbed some data and started building charts right away. I believe it would have been much more helpful to preview the data first, outline some initial questions that I am curious about, and then begin the building process.

References

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1), 3. https://doi.org/10.1186/s41239-020-0177-7

Cortez, P. (2008). Student Performance Data Set. UCI Machine Learning Repository: Student Performance Data Set. Retrieved April 23, 2023, from https://archive.ics.uci.edu/ml/datasets/student+performance

Seshapanpu, J. (2018). Students performance in exams. Kaggle. Retrieved April 23, 2023, from https://www.kaggle.com/datasets/spscientist/students-performance-in-exams