UCD & Fidelity International - Data Analytics Training (July 2022) – Batch C (Ashank Agarwal)

Project Report

GitHub URL

ashank-agarwal/UCDPA_ashankagarwal: GIT Project for CIDAP 22-07-06 [Ashank Agarwal] (github.com)

Abstract

The data used is from a Portuguese secondary school. The facts includes academic and private traits of the scholars as well as very last grades. Predict student performance in secondary education (high school).

Introduction

In the times of pandemic the worst impacted community is of high school students especially in the emerging economies where the alternative education was not so easy set-up. This has been really on back of my mind for a while, to study the trends in academia. I have taken this as an opportunity to pick a student based data to understand how demographics and other soft factors impact the academic performance of students.

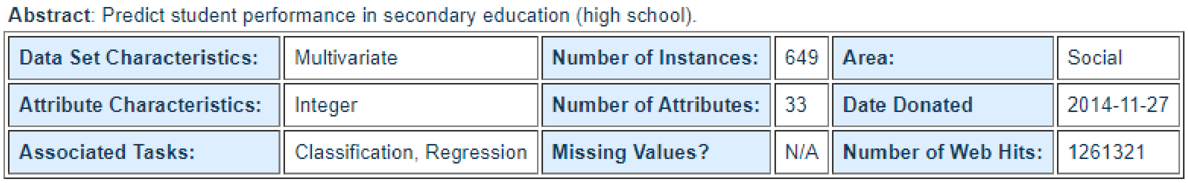
Data insights will be derived using key python libraries used are numpy, pandas, matplotlib, seaborn, re, plotly and sklearn.

We are fully aware that data-based decisions have helped to drive many impactful decisioning and guiding the outcomes. Here students data is used to derive the meaningful insights.

Dataset

Abstract: Predict student performance in secondary education (high school).

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).



Abstract: Predict student performance in secondary education (high school).

Data Set Characteristics:	Multivariate	**Number of Instances:**	649	**Area:**	Social
Attribute Characteristics:	Integer	**Number of Attributes:**	33	**Date Donated**	2014-11-27
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	1261321

(UCI Machine Learning Repository: Student Performance Data Set)

Implementation Process

I have used Jupyter notebook with python interpreter for building the project. I had followed a guided project to build the custom insights for my datasets. There is a well documented tutorial on Kaggle for same and hence I want to call-out that in my submission. You can find all code with output in Git Repo (Link is
already provided in first section)

The analysis has been performed in the below steps using the OSEMN approach –

- Obtain the data
- Scrubbing / Cleaning the data
- Exploring / Visualizing our data
- Modeling the data
- iNterpreting the results

These can be structured further into below 3 broader categories.

1. Identifying the Data Source(s) and Data Collection

The data is identified and saved as csv files. The pandas function were used to read data. This data wsa created as part of research project and selected as our source. Understanding how the data was gathered and what biases, if any, may exist in the data is part of this. Researchers can interpret data more accurately if they understand the data collection process.

2. Exploratory Data Analysis(EDA)

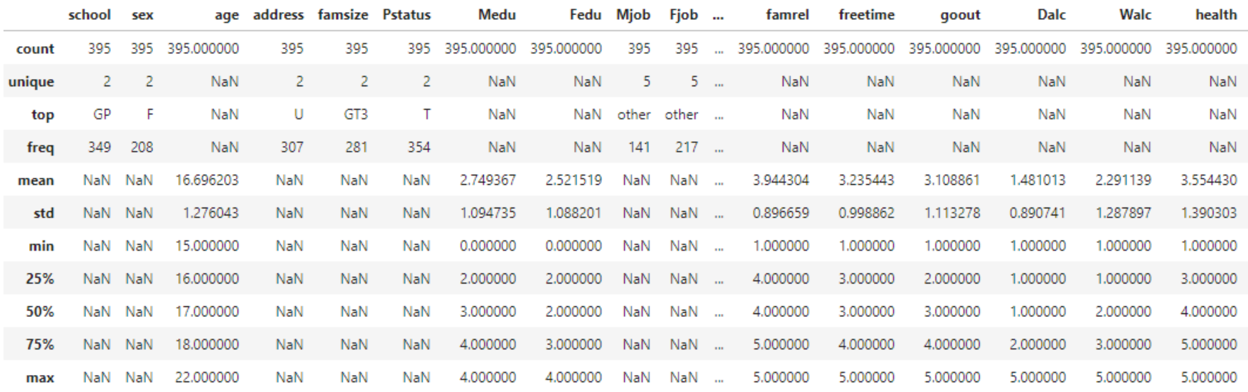
Exploratory data analysis is an important part of working with data. Exploratory data analysis is used to fully understand data and discover all its properties, usually using visual techniques. This allows you to better understand your data and find interesting patterns.

3. Machine Learning

Machine learning is a rapidly growing field of data science with great potential for exploratory data analysis (EDA). EDA has traditionally been done manually by examining patterns and trends in datasets. Machine learning, on the other hand, allows us to automate this process and let computers do the work. There are several machine learning algorithms that can be used with EDA, each with their own strengths and weaknesses.

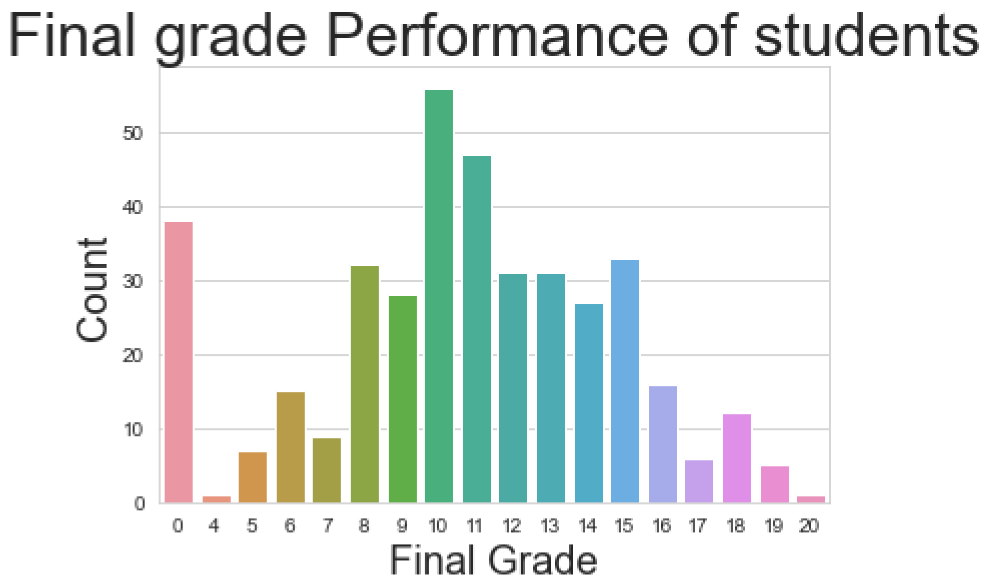
Results & Insights

- **Basic Info of the dataset used.**
    1. Dataset 'df_file' has 395 rows and 33 columns.
    2. No null values found
    3. Features are either categorical or numeric.
    4. Basic Statistics.

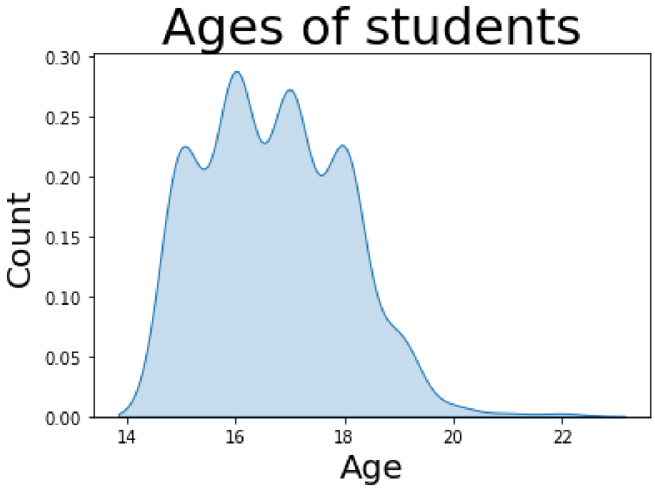


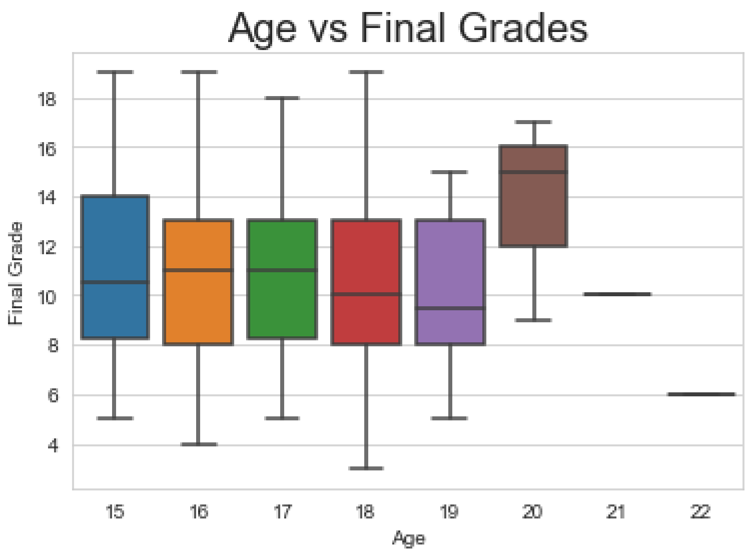
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health
count	395	395	395.000000	395	395	395	395.000000	395.000000	395	395	...	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
unique	2	2	NaN	2	2	2	NaN	NaN	5	5	...	NaN	NaN	NaN	NaN	NaN	NaN
top	GP	F	NaN	U	GT3	T	NaN	NaN	other	other	...	NaN	NaN	NaN	NaN	NaN	NaN
freq	349	208	NaN	307	281	354	NaN	NaN	141	217	...	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	16.696203	NaN	NaN	NaN	2.749367	2.521519	NaN	NaN	...	3.944304	3.235443	3.108861	1.481013	2.291139	3.554430
std	NaN	NaN	1.276043	NaN	NaN	NaN	1.094735	1.088201	NaN	NaN	...	0.896659	0.998862	1.113278	0.890741	1.287897	1.390303
min	NaN	NaN	15.000000	NaN	NaN	NaN	0.000000	0.000000	NaN	NaN	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	NaN	NaN	16.000000	NaN	NaN	NaN	2.000000	2.000000	NaN	NaN	...	4.000000	3.000000	2.000000	1.000000	1.000000	3.000000
50%	NaN	NaN	17.000000	NaN	NaN	NaN	3.000000	2.000000	NaN	NaN	...	4.000000	3.000000	3.000000	1.000000	2.000000	4.000000
75%	NaN	NaN	18.000000	NaN	NaN	NaN	4.000000	3.000000	NaN	NaN	...	5.000000	4.000000	4.000000	2.000000	3.000000	5.000000
max	NaN	NaN	22.000000	NaN	NaN	NaN	4.000000	4.000000	NaN	NaN	...	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

- **Grading distribution of the students**. Apart from the high number of students scoring 0, the distribution is normal as expected. Maybe the value 0 is used in place of null. Or maybe the students who did not appear for the exam, or were not allowed to sit for the exam due to some reason are marked as 0

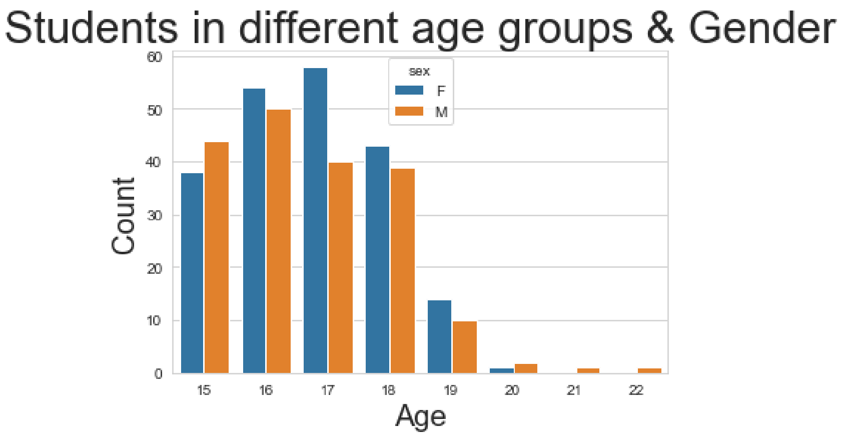


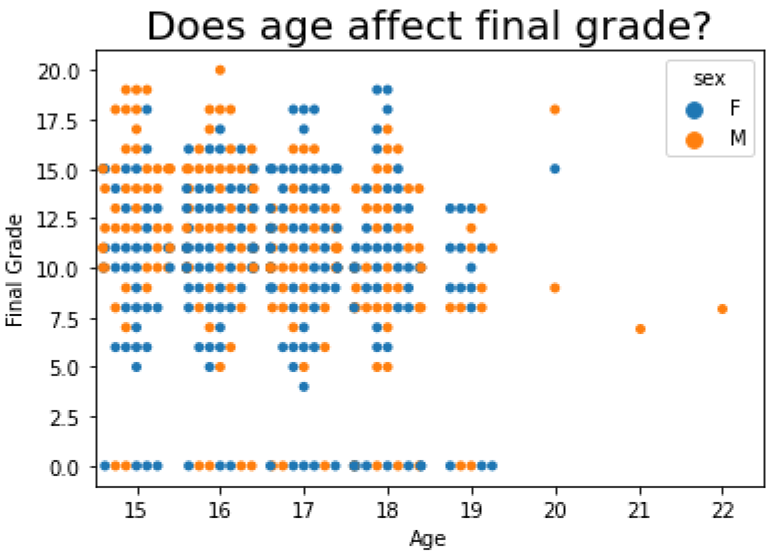
- **Age of students defines their academic performance.**



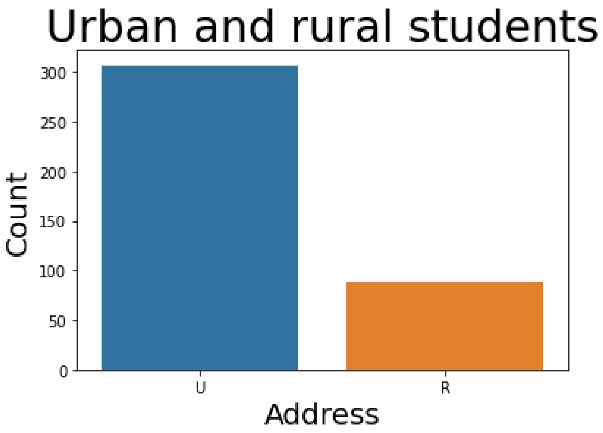


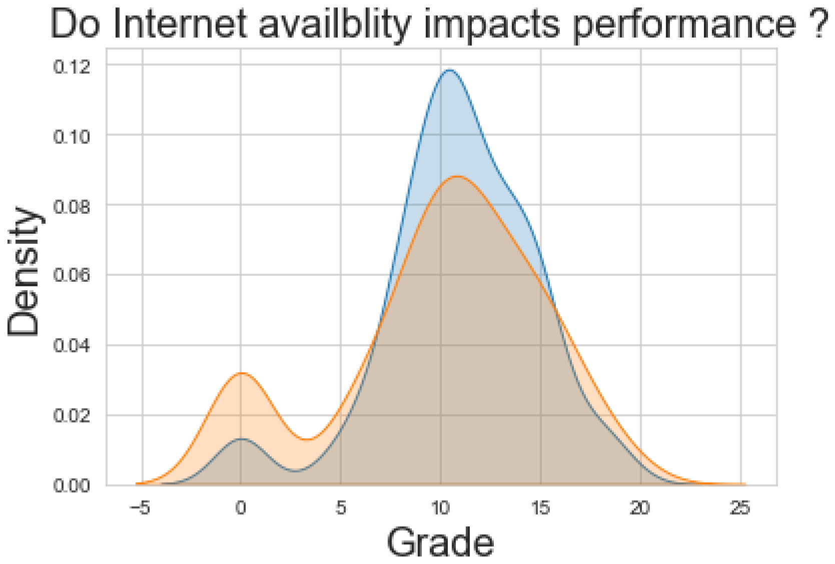
- **Female students do perform better than male counterparts.**

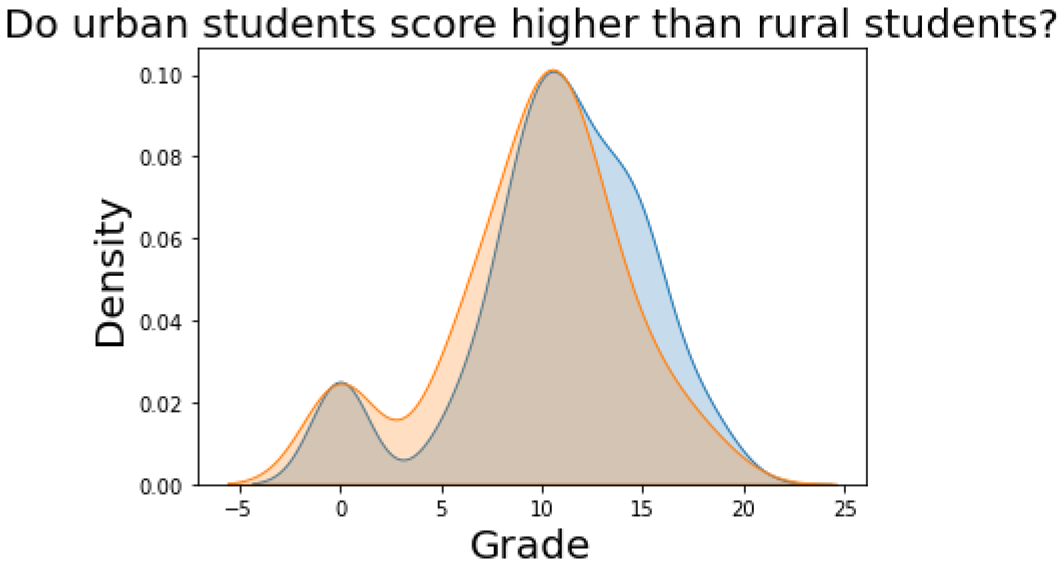




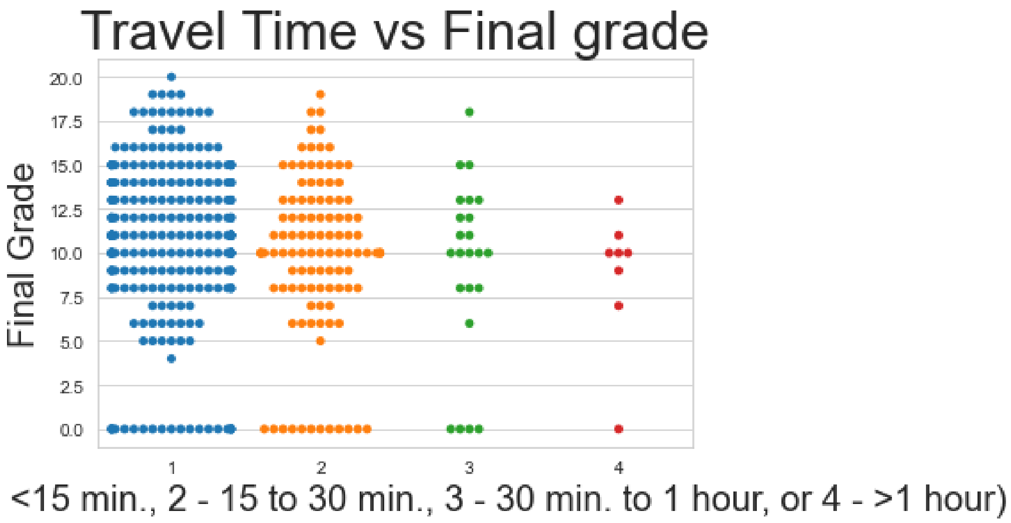
- **Various factors impacting the performance for eg: Urban area, Internet availability etc.**



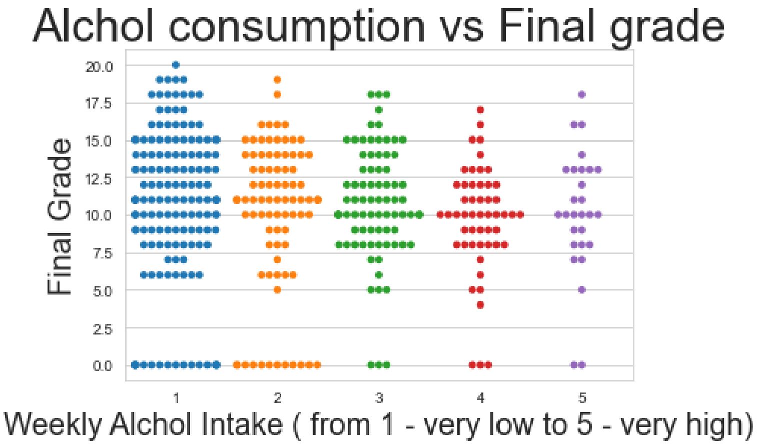




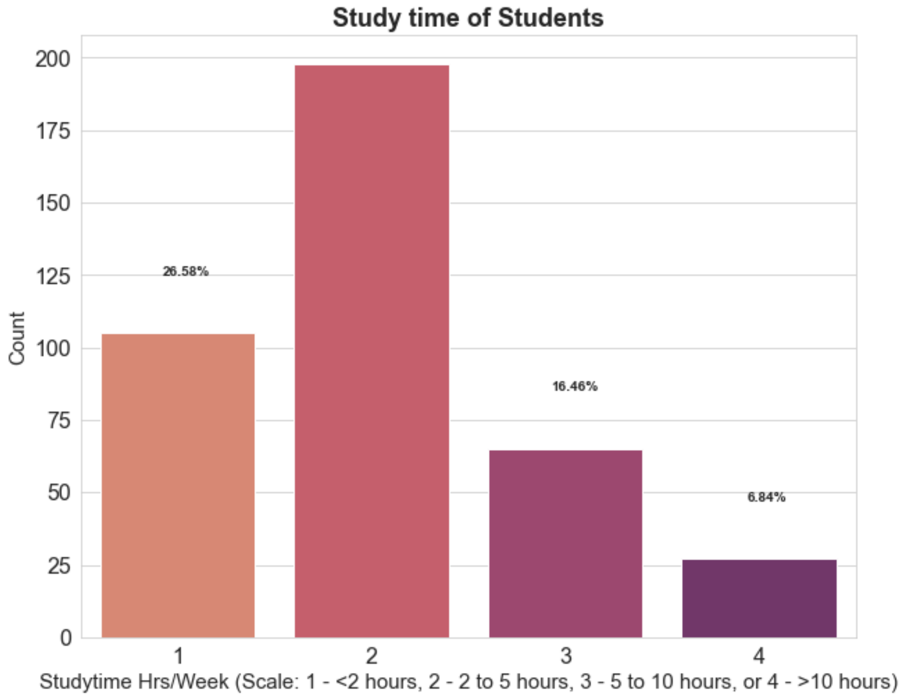
- **Higher Travel time negatively impacts the performance.**



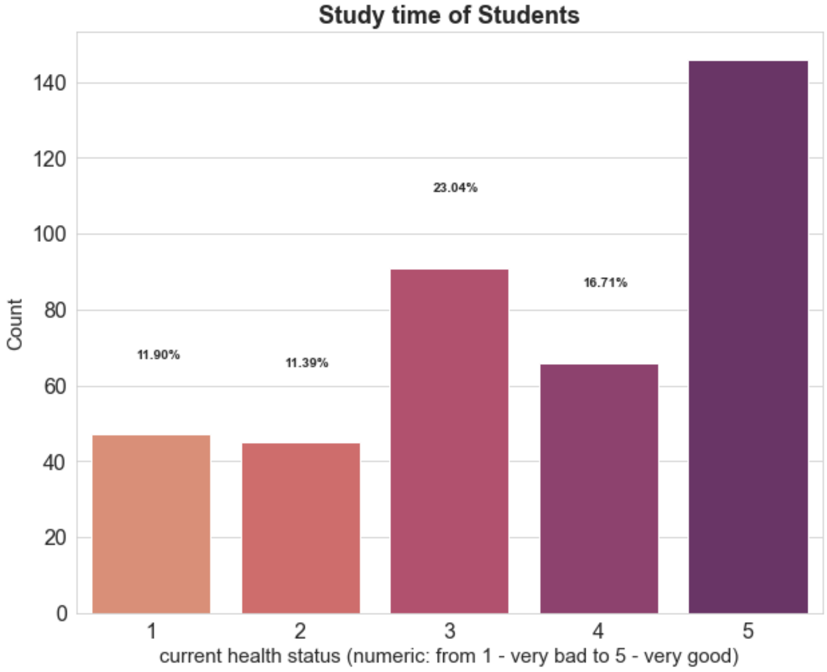
- **Alcohol intake negatively impacts the performance.**



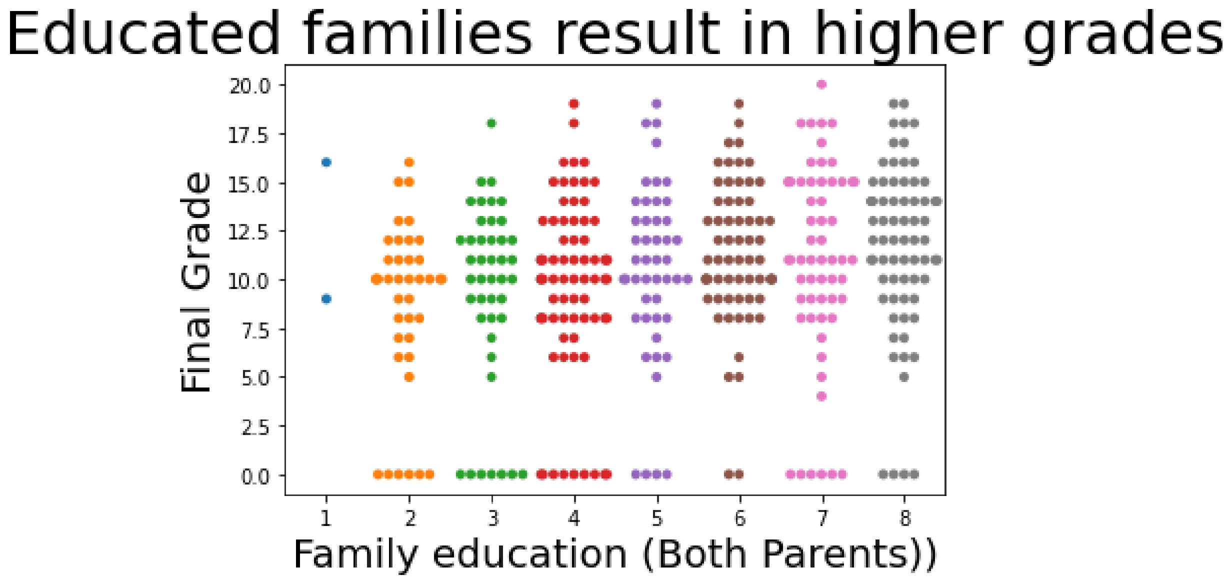
- **Survey data sets shows that maximum of students committed 2 hrs per week.**



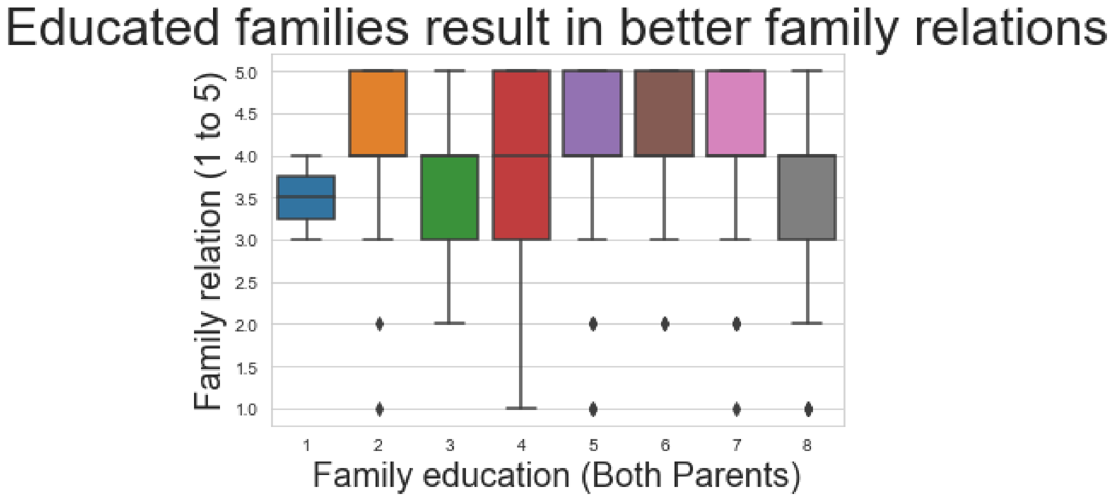
- **Survey data infers that the better health conditions give students better opportunity to devote more time for study.**



- **We can infer from the datasets that education of parents impacts very significantly the performance of students.**

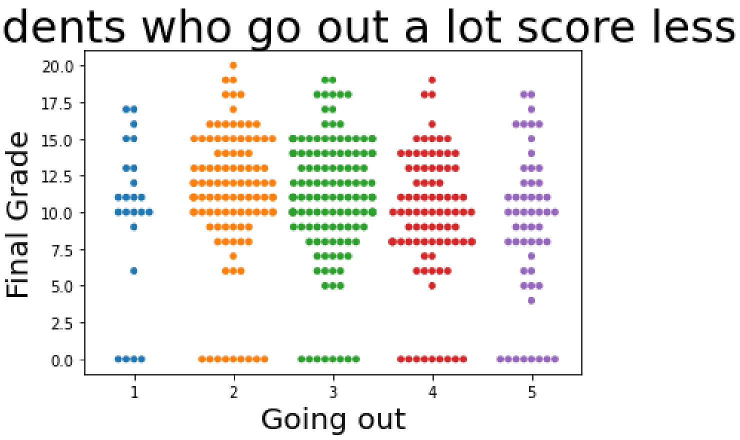


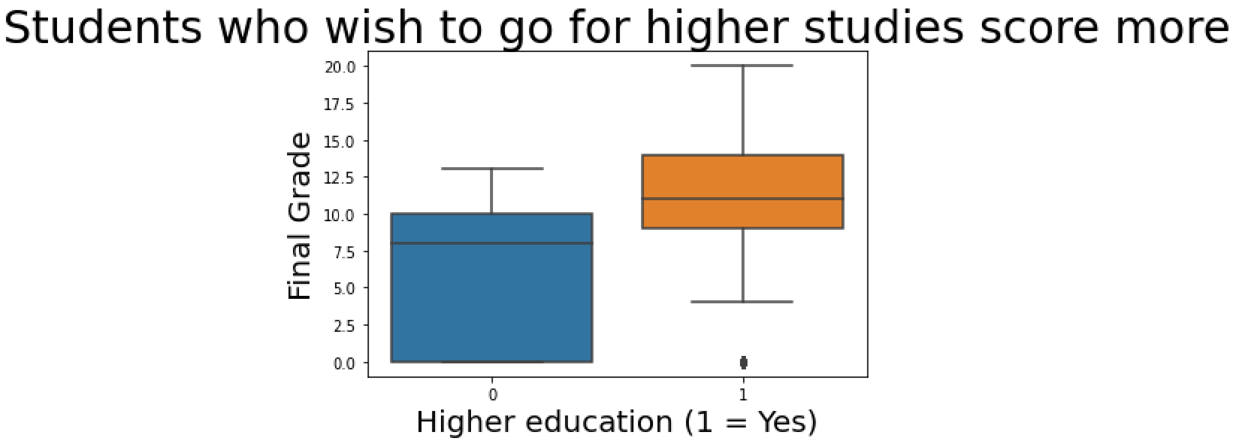
Educated families result in higher grades

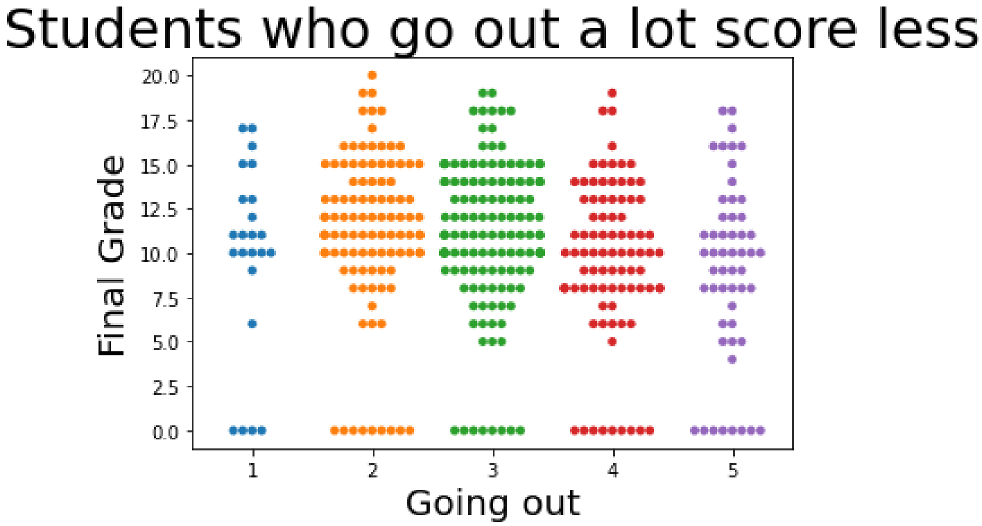


Educated families result in better family relations

- **Other factors like Relationships, future aspirations and going out impacts the performance of students significantly.**

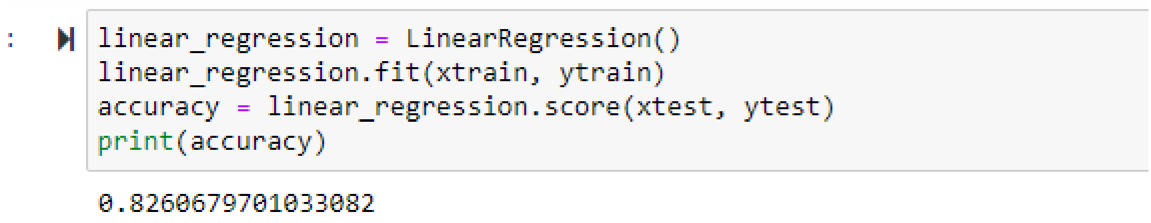






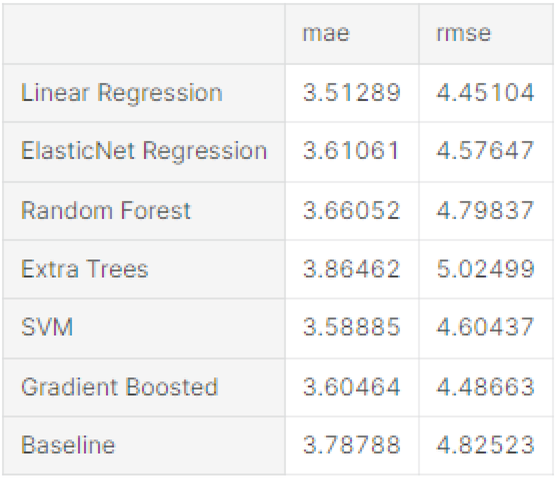
- Machine Learning Outputs :
  We selected the useful columns that we need to train our machine learning model for the task of student grades prediction. Then I declared that the G3 column is our target label and then I split the dataset into 20% testing and 80% training. Now let's see how to train

a linear regression model for the task of student grades prediction:The linear regression model gave an accuracy of about **82%**. Now let’s have a look at the predictions made by the students’ grade prediction model:

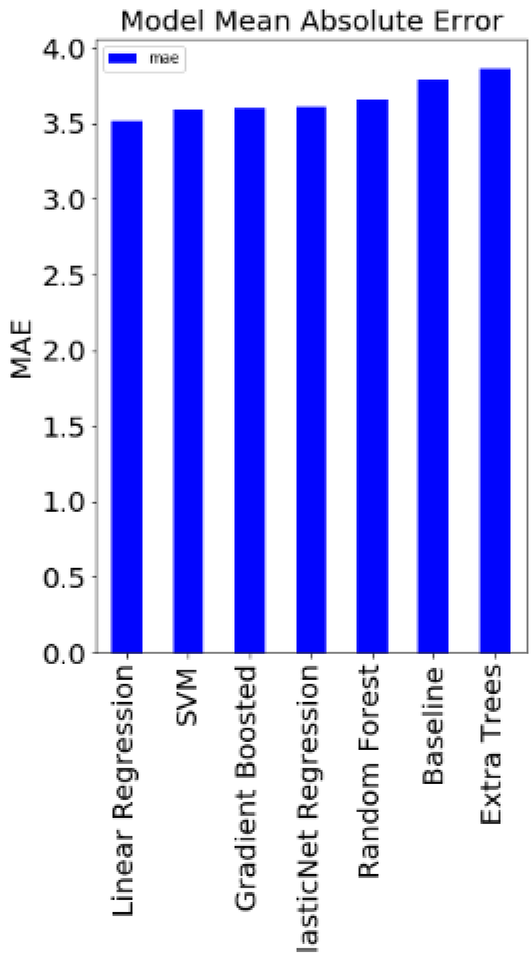


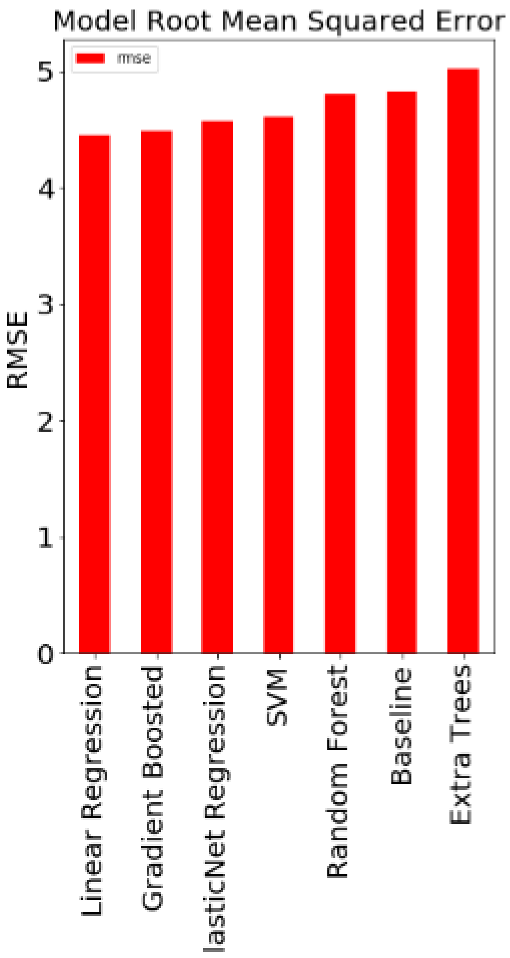
```
linear_regression = LinearRegression()
linear_regression.fit(xtrain, ytrain)
accuracy = linear_regression.score(xtest, ytest)
print(accuracy)
```

```
0.8260679701033082
```



	mae	rmse
Linear Regression	3.51289	4.45104
ElasticNet Regression	3.61061	4.57647
Random Forest	3.66052	4.79837
Extra Trees	3.86462	5.02499
SVM	3.58885	4.60437
Gradient Boosted	3.60464	4.48663
Baseline	3.78788	4.82523





- We have also predicted the used linear prediction model. We have can train a linear regression model for the task of students grade prediction with machine learning using Python (See attached ML_Snippet file in the attached zip)

References

- https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
- https://scikit-learn.org/stable/supervised_learning.html
- https://www.kaggle.com/prashant111/extensive-analysis-eda-fe-modelling
- https://seaborn.pydata.org/generated/seaborn.heatmap.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
- https://www.kaggle.com/code/dipam7/introduction-to-eda-and-machine-learning/n
- Student Grades Prediction with Machine Learning | Aman Kharwal (thecleverprogrammer.com)