COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan

Process Description

1. Clean
    - Use pandas in python to transform data to disregard outlier data.
    - Scripts Utilized: **clean_department_table.py**, **clean_counciling_table.py**, **clean_employee_table.py**, **clean_performance_table.py**, enhanced by using **cleanAll.py** which will clean all the data lowering overhead.
2. Load
    - Create a database in MariaDB with the described ER diagram in the assignment specifications.
    - Load in 'department.csv', 'employee.csv', 'student.csv', and 'performance.csv' using a script **'loadAll.sql'** which executes '**load_student_csv'**. ' '**load_department_csv.sql', 'load_employee_csv.sql', 'load_performance_csv'.** We utilized load all since key dependencies require tables to be set up in order, load all ensures order is maintained.
    - Maintenance scripts **'reset_tables.sql'** used to reset tables and remove all keys to recreate and retest our loading.
3. Transform and validate
    - Use python scripts to ensure that the data is valid and usable.
    - If the data is unusable, exceptions are reported and listed below in this document.
    - The programs output a clean version of the data, along with any data that has been filtered out due to exceptions
4. Extract
    - Join the 4 tables to create a comprehensive aggregate table. At this point, any inconsistencies are removed. **'Join.sql**' is utilized to match dependencies to create one aggregate table "**Comprehensive Data**"
    - Use **'export.sql'** to export comprehensive data for analysis.
        - VM Constraint -> Had to output export to /tmp/ due to permission restraints. Need to move to the output directory and chown to self as a temporary fix.
        - DBeaver -> Half of our team used their own instances of MariaDB with a RDMS which allowed simple export to drive.
    - Load transformed validated aggregated data into Jupyter notebook for descriptive and predictive data analytics.

Setup Instructions

1. Download
2. Change user from pgill914 to yours in file paths
3. In Linux terminal: "/cleanLoad/cleanScripts/cleanAll.sql"
4. Go into MySQl and "create database student_performance"
5. In mySQL terminal: "source path/to/reset_tables.sql"

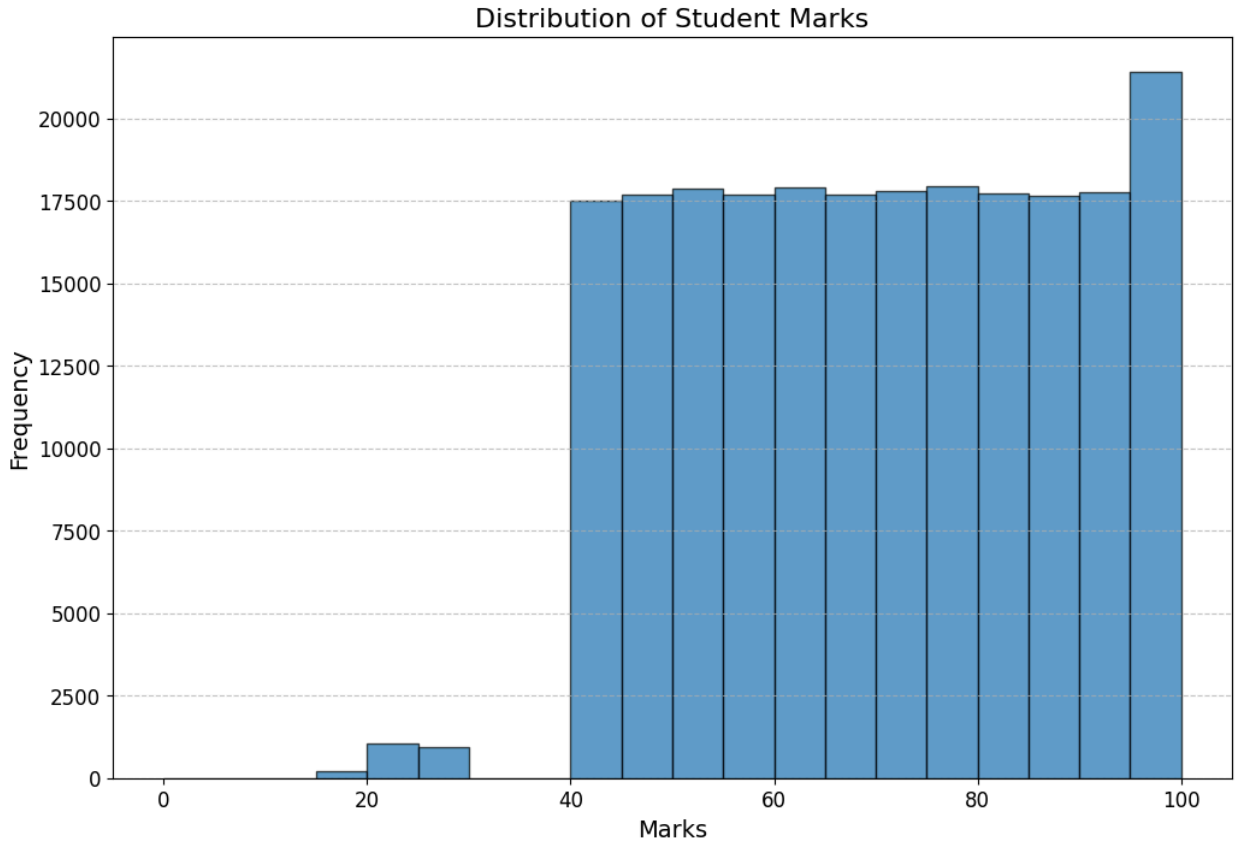
COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan

6. In mySQL terminal: "source path/to/loadAll.sql"
7. Create joined table In mySQL terminal: "source path/to/join.sql"
8. Export table in mySQL terminal: "source path/to/export.sql"
9. Move comprehensive_data.csv from tmp from linux command line "sudo mv /tmp/comprehensive_data.csv /path/to/visualization/
10. Launch jupyter and run all cells

Data Mining Analysis

Descriptive
With our given data we decided to look at the data and make some charts on what we felt was important to see.

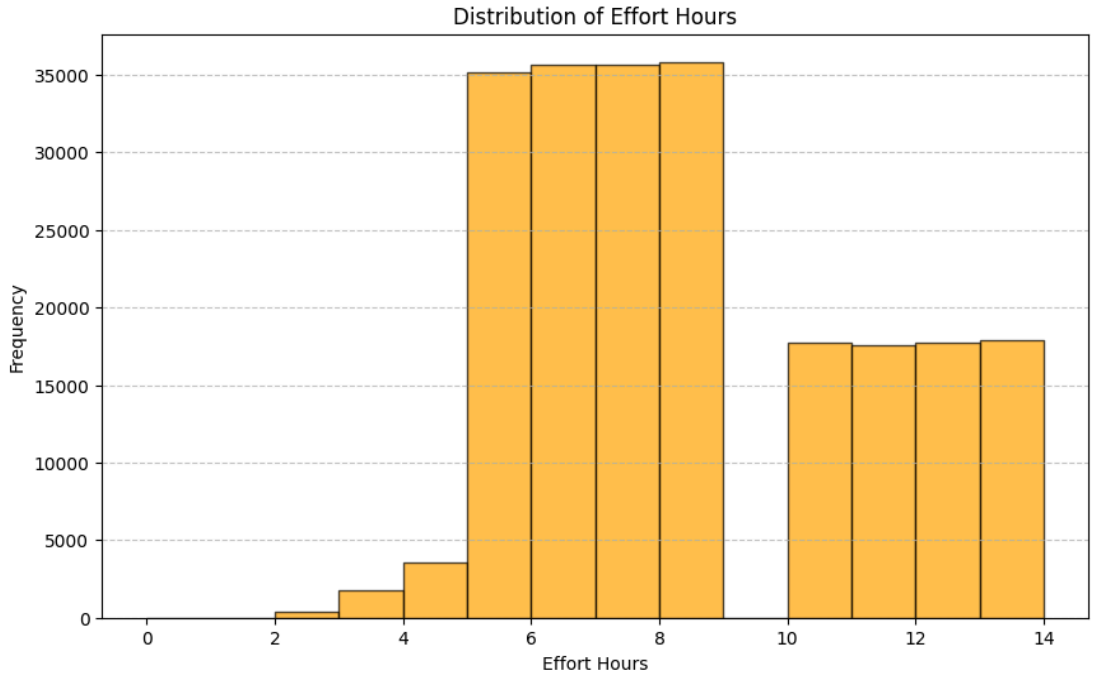
We first analyzed the distribution of student marks. The chart shows that the majority of students scored above 40, indicating a strong overall performance. Notably, a significant number of students achieved grades in the range of 90 to 100, highlighting a cluster of high achievers.



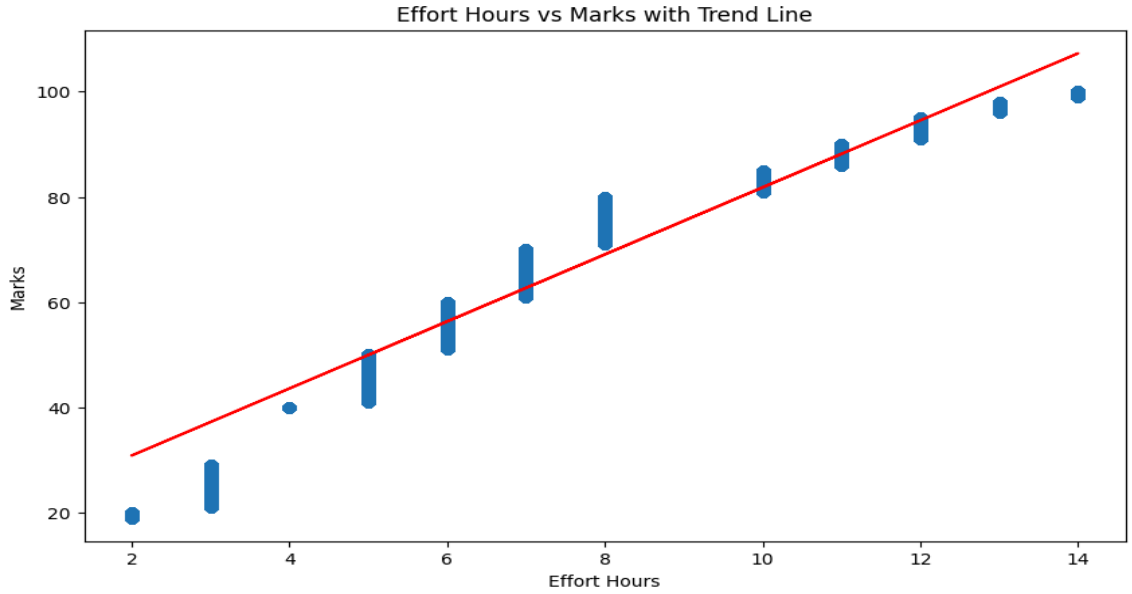
The chart illustrates the distribution of effort hours spent by students per paper. The majority of students invested between 5 and 8 hours on a paper, with no apparent instances of exactly 9 hours. A secondary group of students dedicated between 10 and 14 hours. The absence of data for 9 hours suggests that students might either stop at 8 hours or, if continuing, commit at least

COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan

2 additional hours, resulting in 10 hours or more.



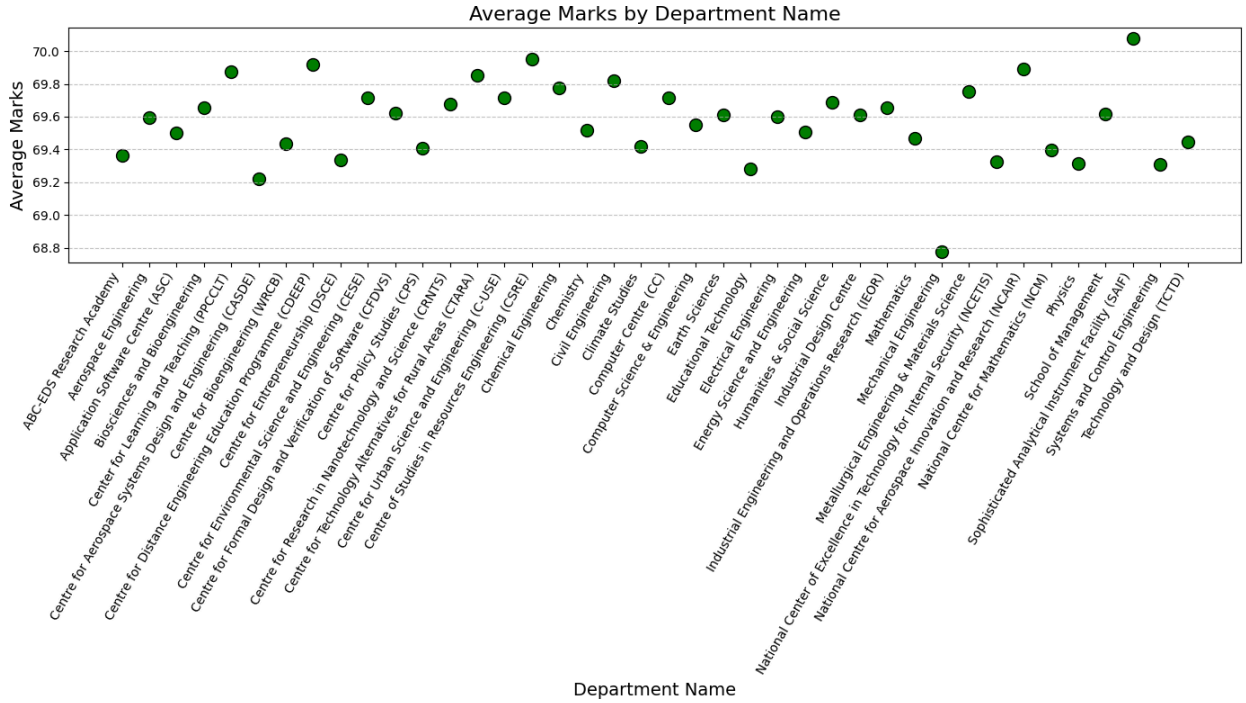
The chart below depicts the relationship between effort hours and marks, including a trend line to illustrate the overall trend. It is evident that there is a positive correlation between the number of hours spent studying and the marks achieved. As students put in more effort hours, their marks tend to increase, as shown by the upward sloping red line trend. This suggests that increased study time generally results in better performance, highlighting the importance of dedicated effort in achieving higher grades.



COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan

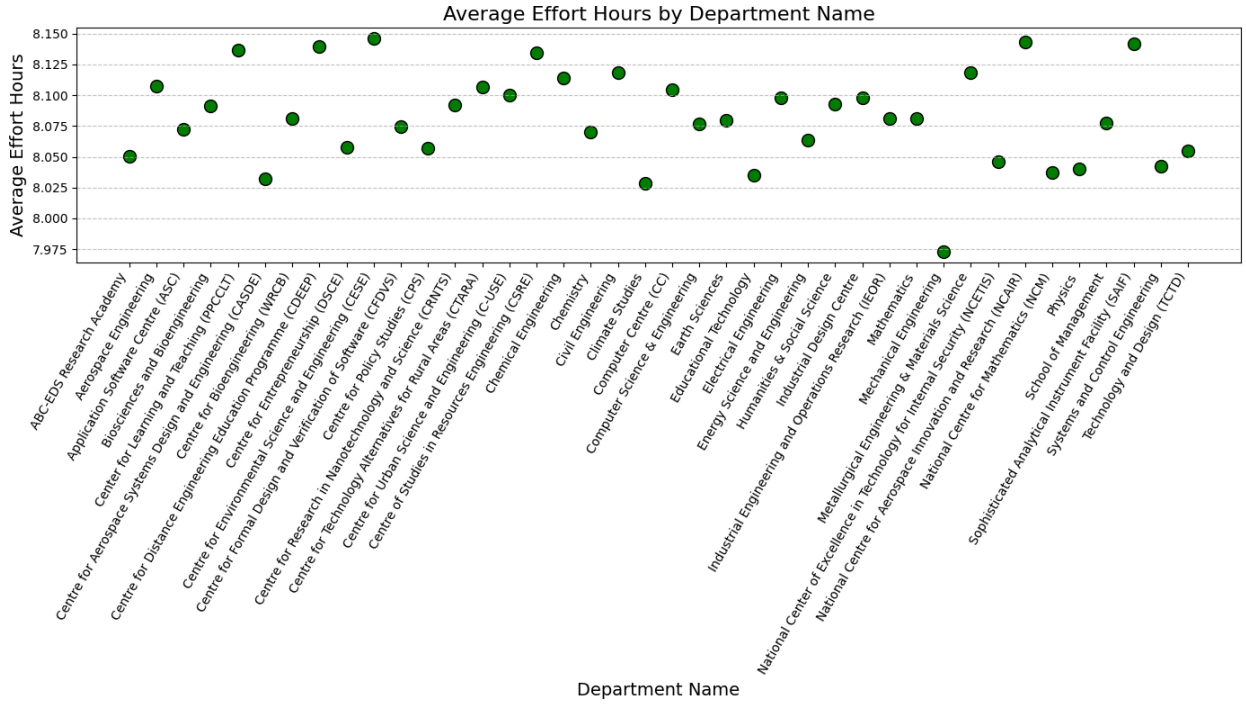
For fun, we decided to see which departments had the highest and lowest marks and effort hours.

The chart displays the average marks achieved by students across various departments. Most departments have average marks between 69.0 and 69.8 with a few departments exceeding 70. The results indicate that student performance is fairly consistent across departments with slight variations in average marks. The department with the lowest average was Mechanical Engineering at 68.8 while the highest was Sophisticated Analytical Instrument Facility at 70.08.



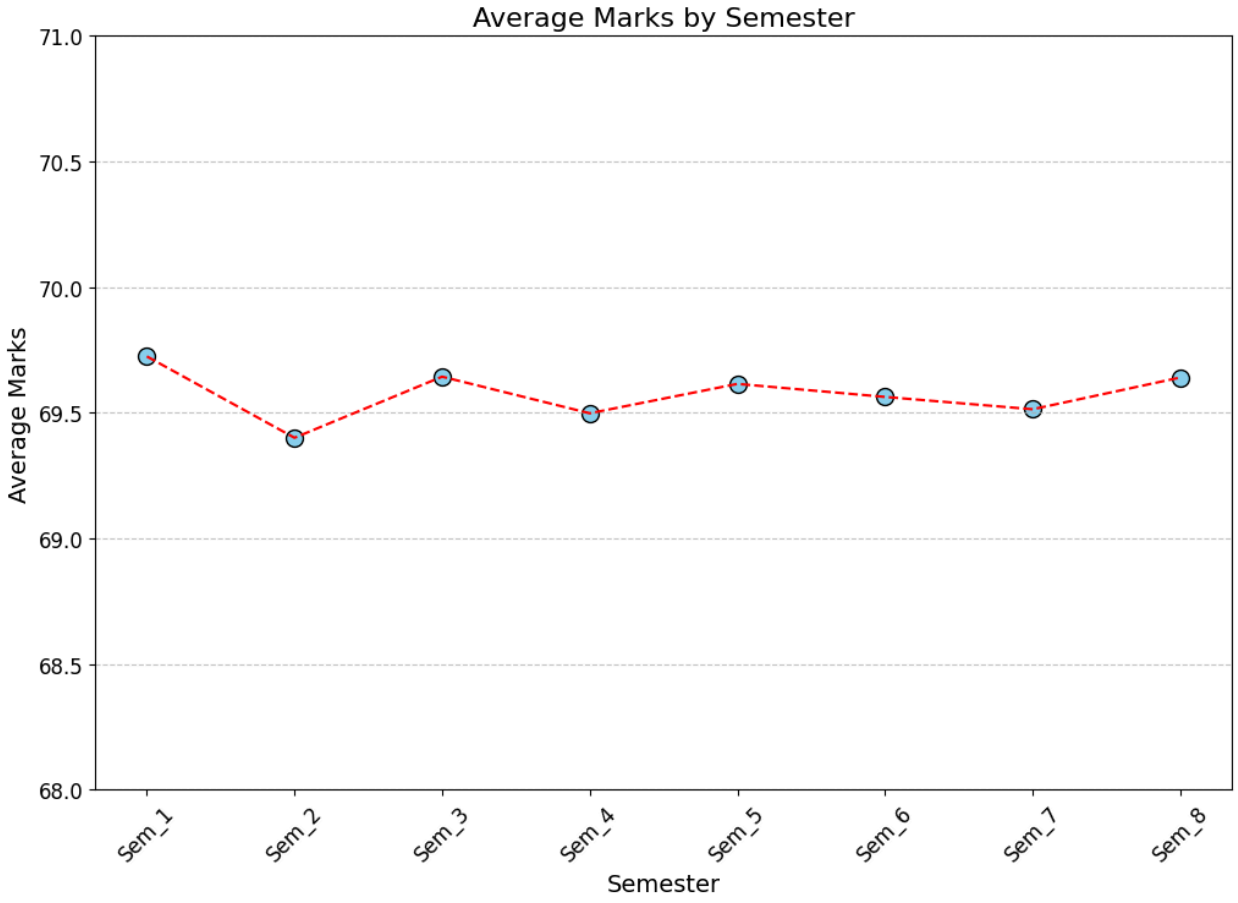
The next chart is then the average effort hours by students across different departments. The average effort hours ranges between 7.97 and 8.15 hours, indicating a fairly consistent level of study effort across all departments. There are slight variations among the departments, but overall, most departments cluster around similar values of average effort hours. This suggests that, regardless of the field of study, students are dedicating a comparable amount of time to their coursework, with only minor deviations. The highest amount of hours being 8.15 by Centre for Environmental Science and Engineering and the lowest being 7.97 hours by Mechanical Engineering. There could be correlation between Mechanical Engineering having the lowest average mark due to having the lowest amount of hours of effort.

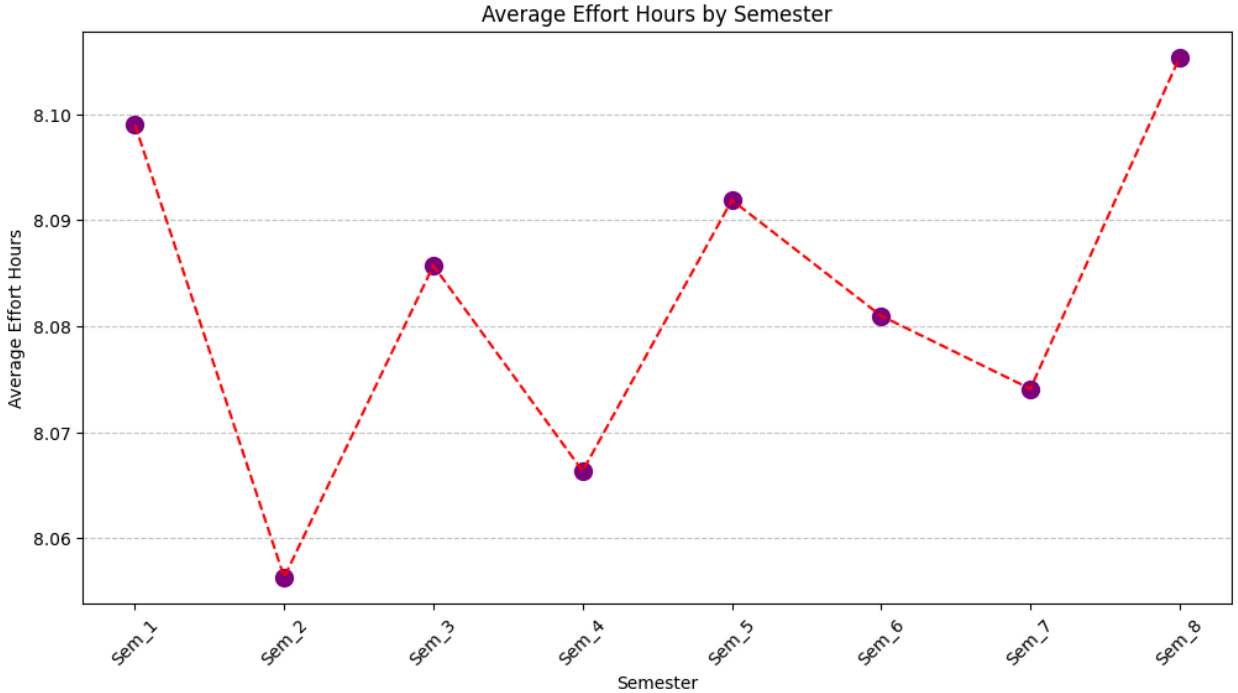
COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan



Our final two charts are the average marks and effort hours across each semester. Overall, the average marks fluctuate slightly but remain relatively stable between 69.0 and 70.0 across all semesters. The effort hours however show a little more variability as some semesters seem to require more effort hours than others. Notably, semester 1 and semester 8 have the highest average effort hours while semester 2 has the lowest. It seems that students tend to not care about semester two as much as the other semesters as it seems that effort hours are the lowest causing average marks to be the lowest as well in semester 2. It could be that students are fresh and want to try their hardest in the first semester which causes the increase in effort hours while at the end, students are looking to finish strong in their last semester. The higher average grade in the first semester could be seen being caused by the easier content compared to the last semester. Semester 2 could be the lowest due to students feeling like the content is easy, thus putting less effort which then affects the marks. The low mark can act as a wakeup call to put in more effort hours to get better marks as seen in the increase in average hours.

COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan





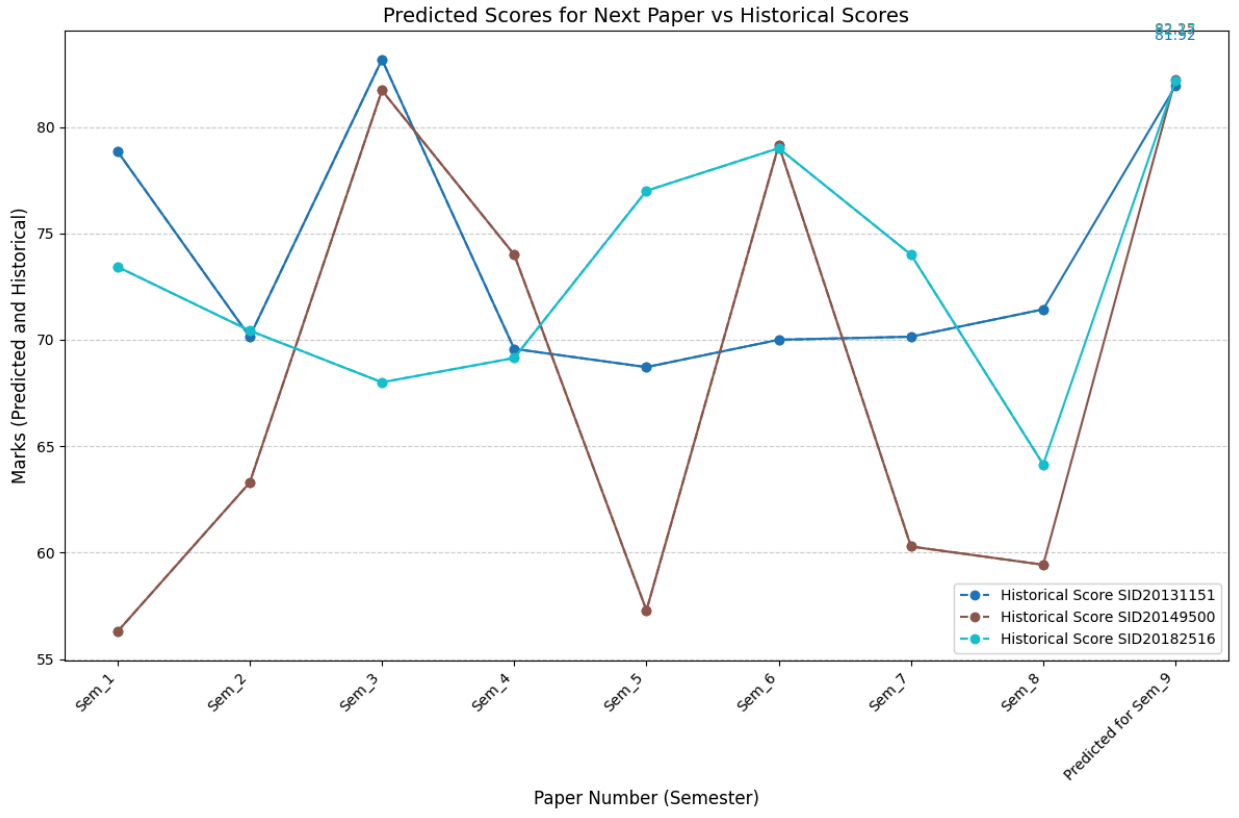
COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan

Predictive

With our given data we maintained Student_ID, Department_Admission, Department_Name, Semester_Name, Paper_ID, Marks, Effort_Hours. We hypothesized from our personal university experience that the effort put into a paper will result in an increased mark, to confirm and test this we created a regression model which analyzes the relationship between effort hours and marks. We have historical data for our students which allows us to train our model based on the time variance data from past semesters and papers. We are limiting our model to only predict for three students, this speeds up our analytics and we should be able to see consistent predictions for the students as they are in the same academic school.

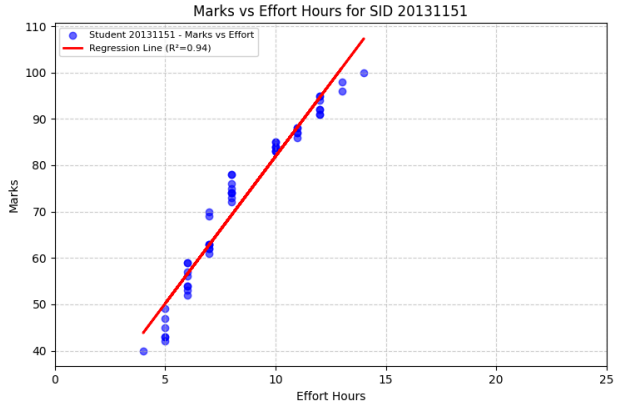
Predicted Score for next paper:

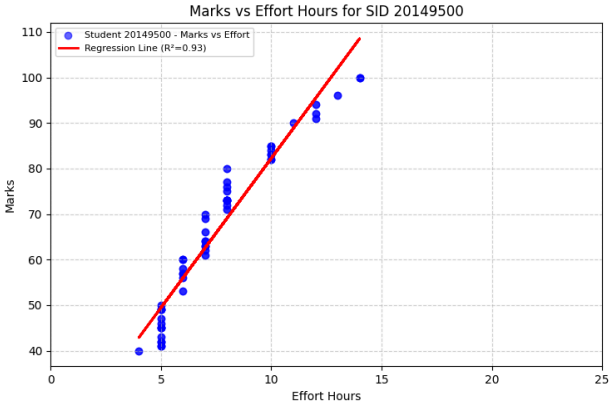
Student	Predicted Score in next paper	Department
SID20131151	81.92/100	IDEP6347
SID20149500	82.25/100	IDEP4308
SID20182516	82.17/100	IDEP3062

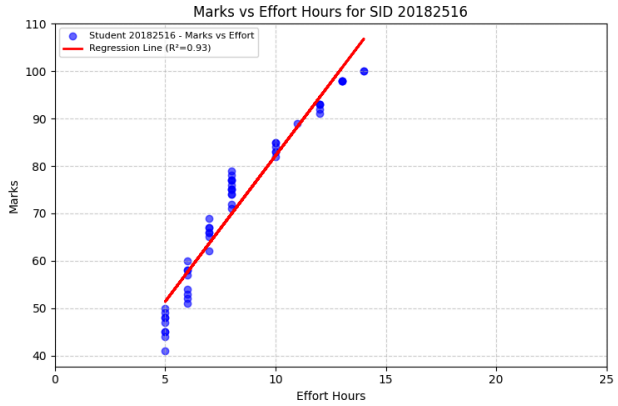


COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan

Results from Effort vs Marks:







Our model with its training was able to display a clear trend that the more effort you put in the more marks you will get, which was in line with our hypothesis. Calculating R^2 gives us values ranging from 94-95% which are extremely high values which confirm that effort hours are a strong predictor of marks.

Reporting

To ensure data quality, we implemented a comprehensive pre-processing step to work exclusively with high-quality, complete data. During this process, any rows containing exceptions such as missing values, outliers, or inconsistencies are identified, removed, and logged in a separate CSV file for each dataset. These invalid data files provide a record of data exclusions for error checking. The cleaned data is then saved in new CSV files, which will optimize the computations in the following steps. The entire pre-processing is handled using pandas in python.

Department_Information: Exceptions were thrown for the following reasons:

- Attributes ‘Department_ID’ or ‘Department_Name’ are not unique.
- Attribute ‘DOE’ is less than the year 1900.
- Any attribute is missing values.

COMP 4522 Assignment 2 Report
Group: Ben Cacic, Stanley Chow, Puneet Gill, Andrew Phan

```
Department_ID,Department_Name,DOE
IDEPT5528,Sanitation and Digital Gaming,
IDEPT9009,Laser Technology Enhancements,
IDEPT9999,Centre for Studies of Mars Ecology,13/03/2025
IDEPT5528,Biosciences and Bioengineering,6/28/1943
IDEPT1825,Mechanical Engineering,9/21/1971
IDEPT3868,Center for Learning and Teaching (PPCCLT),3/26/1982
IDEPT5528,Sanitation and Digital Gaming,
IDEPT7005,Centre of Studies in Resources Engineering (CSRE),8/22/1966
IDEPT7005,Centre of Studies in Craft Engineering (CSCE),8/22/1966
IDEPT9009,Centre for the Study of Ecology in Mars,7/9/2025
IDEPT3868,Center for Learning and Teaching (PPCCLT),3/26/1982
IDEPT9009,Laser Technology Enhancements,
IDEPT1825,Materials Strength Testing,9/21/1971
```

Student_Counceling_Information: Exceptions were thrown for the following reasons:

- Missing values on the attribute ‘Department_Admission’.
- Attribute ‘Department_Admission’ does not exist.

```
Student_ID,DOA,DOB,Department_Choices,Department_Admission
SID20135073,7/1/2013,12/7/1995,,
```

Student_performance_Data: Exceptions were thrown for the following reasons:

- Attribute ‘Marks’ holds values that are not between 0 to 100.
- Attribute ‘Hours’ was a negative value.
- A ‘Student_ID’ has more than one mark per each ‘Paper_ID’.
- Any attribute is missing values.

```
Student_ID,Semster_Name,Paper_ID,Paper_Name,Marks,Effort_Hours
SID20131189,Sem_1,SEMI0015910,Paper 4,-49.0,0.0
SID20131191,Sem_5,SEMI0055015,Paper 6,207.0,14.0
SID20131231,Sem_1,SEMI0016208,Paper 5,-100.0,14.0
SID20131303,Sem_3,SEMI0031818,Paper 4,140.0,14.0
SID20179280,Sem_4,SEMI0044518,Paper 6,,
SID20182774,Sem_8,SEMI0086600,Paper 6,999.0,5.0
SID20189989,Sem_6,SEMI0064181,Paper 4,,6.0
SID20147406,Sem_6,SEMI0067259,Paper 2,78.0,-3.0
```