Databases 2 - Assignment 2 - Report

Seila, Kiera, James, Diesel

ETL Tasks

- We extracted the data into the tables
- We then cleaned and validated the data according to the table below

We then performed the predictive and descriptive analysis.

Design Decisions

We opted to use SQLite for its simplicity over MariaDB. The entire process of extracting the data from the CSV files, adding to the database, cleaning, and mining was done in our Python Jupyter Notebook. Data is extracted from the CSV files directly into the SQLite database. The schema was modified from the one specified in the assignment description to allow for and inconsistencies, duplicates and other issues that were prevalent in the data. Additionally, an index (id column) was added to ensure all rows had a unique primary key, regardless of duplicates.

Data cleaning was done in Pandas from data that was extracted from the SQLite database. Data was cleaned as per the requirements table, with each requirement generating a table with the rows containing errors. These tables are provided below. Exception reporting is done on the original data, rather than the previously cleaned data as to show rows that may have issues in multiple areas. Cleaned data is then joined in Pandas in a similar fashion to a SQL inner join, then filtered and used for data mining.

Cleaned and Validated as per this table

Table / Relation	Attribute	Validation Required	Action
Department_Information	Department_ID	Uniqueness	Report Exception
Department_Information	Department_Name	Uniqueness	Report Exception
Department_Information	DOE	Year >= 1900	Report Exception
Department_Information	All	Missing values	Report Exception
Employee_Infomation	None	None	None
Student_Counceling_Information	Department_Admission	Missing Values	Report Exception

Table / Relation	Attribute	Validation Required	Action
Student_Counceling_Information	Department_Admission	Department_Admission does not exist	Report Exception
Student_Performance_Data	Marks	Range: 0 to 100	Discard entries with issues, and report them
Student_Performance_Data	Hours	Min = 0 Max = any positive integer	Discard entries with issues, and report them
Student_Performance_Data	Student_ID and Paper_ID	A given Student_ID cannot have more than 1 mark per each Paper_ID	Report Exception
Student_Performance_Data	All	Missing values	Discard entries with issues, and report them

Data Inspection, Exception Reporting and Cleansing

Cleaning and validating Department Information data

ISSUE: Non Unique Department_ID's

id	Department_ID	Department_Name	DOE
11	IDEPT1825	Mechanical Engineering	9/21/1971
45	IDEPT1825	Materials Strength Testing	9/21/1971
15	IDEPT3868	Center for Learning and Teaching (PPCCLT)	3/26/1982
35	IDEPT3868	Center for Learning and Teaching (PPCCLT)	3/26/1982
1	IDEPT5528	Biosciences and Bioengineering	6/28/1943
21	IDEPT5528	Sanitation and Digital Gaming	None
24	IDEPT7005	Centre of Studies in Resources Engineering (CSRE)	8/22/1966
25	IDEPT7005	Centre of Studies in Craft Engineering (CSCE)	8/22/1966
27	IDEPT9009	Centre for the Study of Ecology in Mars	7/9/2025
39	IDEPT9009	Laser Technology Enhancements	None

ISSUE: Non Unique Department_Name's

id	Department_ID	Department_Name	DOE
15	IDEPT3868	Center for Learning and Teaching (PPCCLT)	3/26/1982
35	IDEPT3868	Center for Learning and Teaching (PPCCLT)	3/26/1982

ISSUE: Invalid DOE, or DOE < 1900

id	Department_ID	Department_Name	DOE
21	IDEPT5528	Sanitation and Digital Gaming	None
31	IDEPT1677	ABC-EDS Research Academy	7/10/1849
39	IDEPT9009	Laser Technology Enhancements	None
42	IDEPT9999	Centre for Studies of Mars Ecology	13/03/2025

ISSUE: NULL/NaN Values

id	Department_ID	Department_Name	DOE
21	IDEPT5528	Sanitation and Digital Gaming	None
39	IDEPT9009	Laser Technology Enhancements	None

Cleaning and validating Student Counselling data

ISSUE: NULL/NaN Values

id	Student_ID	DOA	DOB	Department_Choices	Department_Admission
298	SID20135073	7/1/2013	12/7/1995	None	None

ISSUE: Missing foreign keys in deptInfo for Department_Admission

id	Student_ID	DOA	DOB	Department_Choices	Department_Admission

(Empty, no issues found)

Cleaning and validating studentPerformance data

ISSUE: Out of range Mark values

id	Student_ID	Semster_ Name	Paper_ID	Paper_ Name	Marks	Effort_ Hours
328	SID20131189	Sem_1	SEMI0015910	Paper 4	-49	0
414	SID20131191	Sem_5	SEMI0055015	Paper 6	207	14
551	SID20131231	Sem_1	SEMI0016208	Paper 5	-100	14
840	SID20131303	Sem_3	SEMI0031818	Paper 4	140	14
181488	SID20182774	Sem_8	SEMI0086600	Paper 6	999	5

ISSUE: Negative Effort_Hours values

id	Student_ID	Semster_ Name	Paper_ID	Paper_ Name	Marks	Effort_ Hours
59635	SID20147406	Sem_6	SEMI0067259	Paper 2	78	-3

ISSUE: Duplicate mark entries for a student and a paper

id	Student_ID	Semster_ Name	Paper_ID	Paper_ Name	Marks	Effort_ Hours

(Empty, no issues found)

ISSUE: NULL/NaN Values

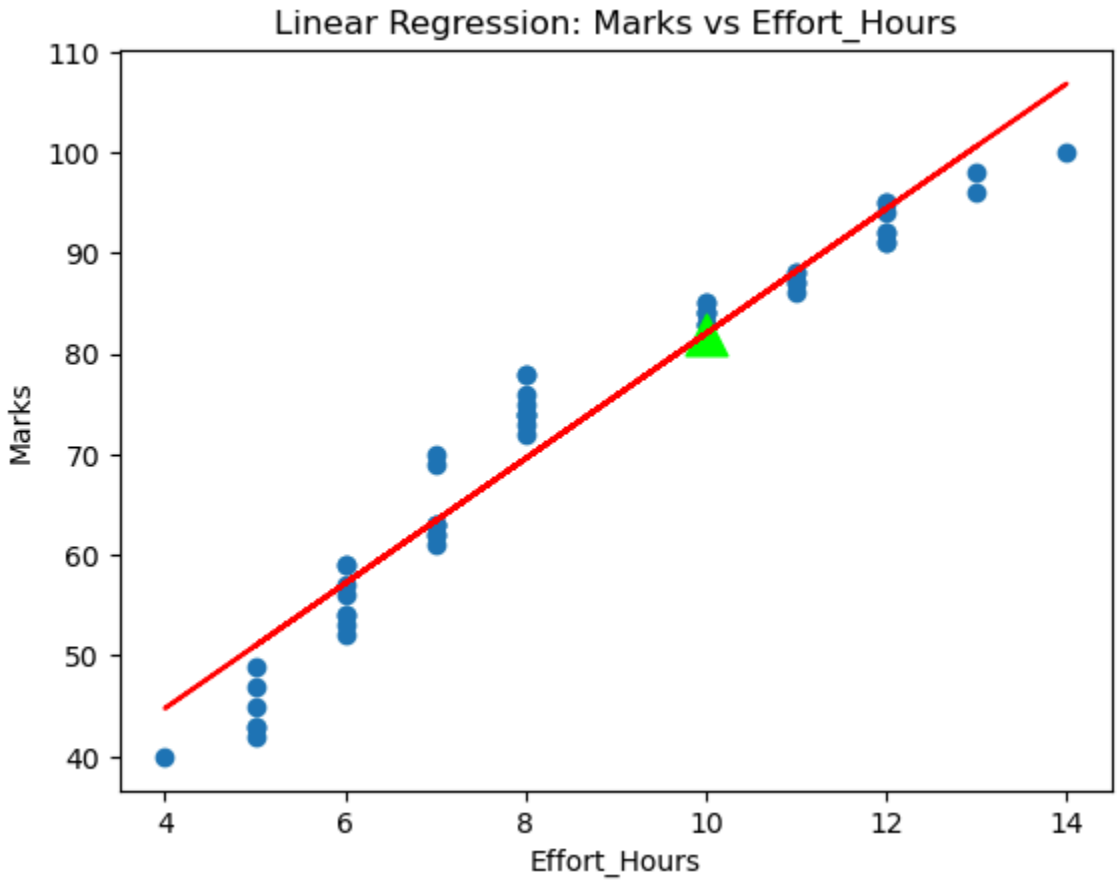
id	Student_ID	Semster_ Name	Paper_ID	Paper_ Name	Marks	Effort_ Hours
125	SID20131171	Sem_3	SEMI0031818	None	87	11
172218	SID20179280	Sem_4	SEMI0044518	Paper 6	nan	nan
209593	SID20189989	Sem_6	SEMI0064181	Paper 4	nan	6

Data Mining

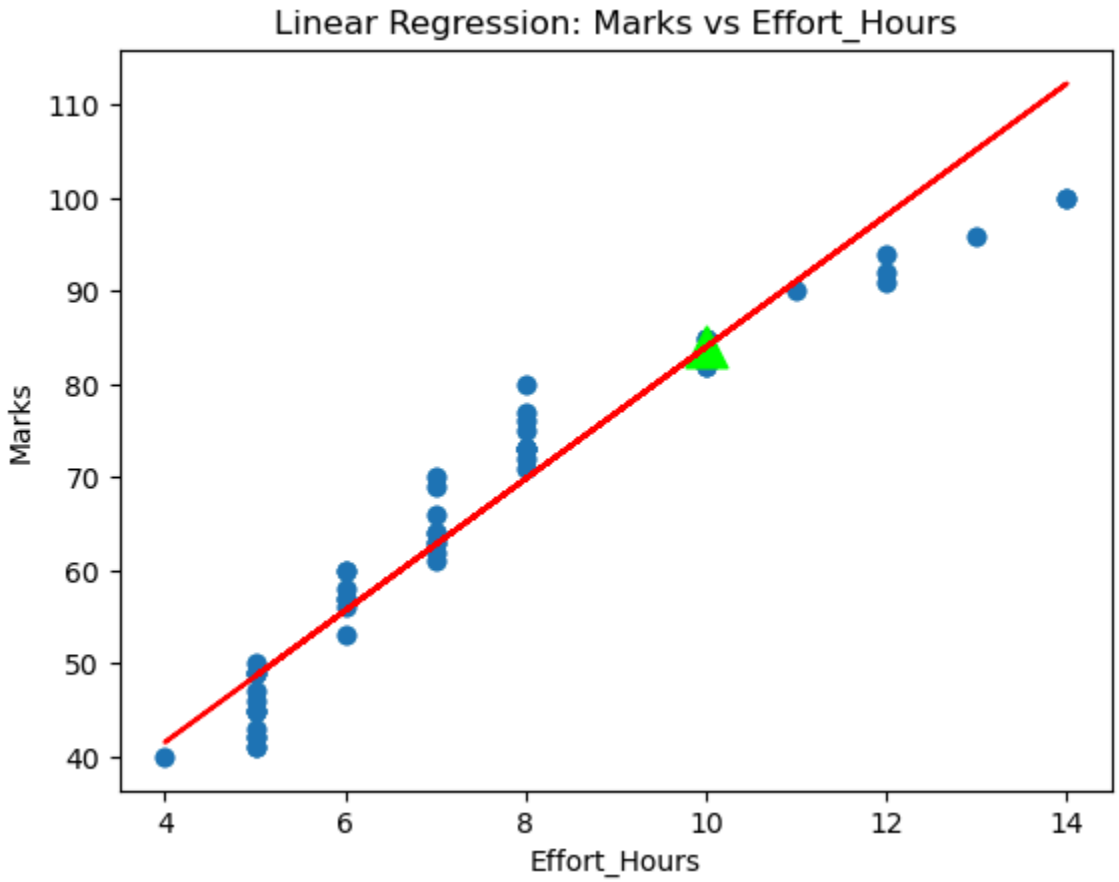
Student Marks/Effort_Hours prediction

Student	Predicted Score in next paper	Department
SID20131151	81.99%	IDEPT6347
SID20149500	83.99%	IDEPT4308
SID20182516	82.40%	IDEPT3062

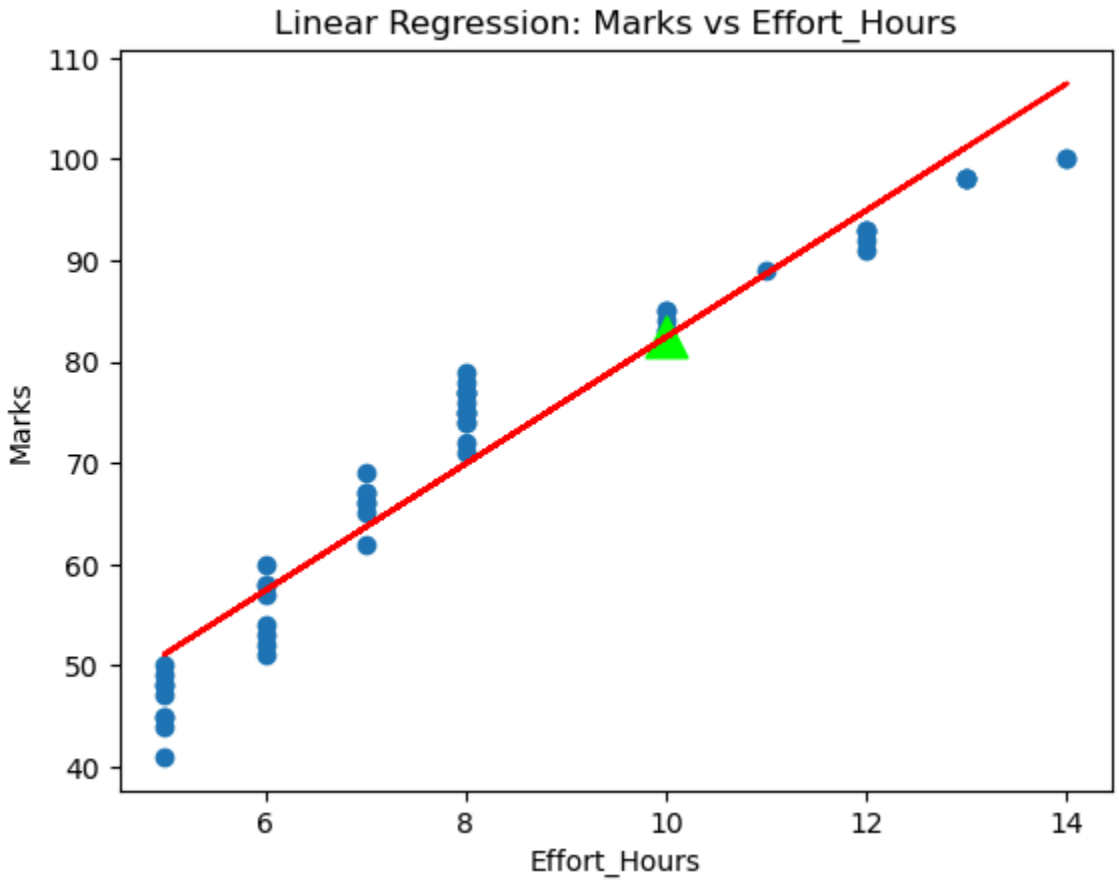
Linear Regression for Student 1 (SID20131151)



Linear Regression for Student 2 (SID20149500)



Linear Regression for Student 3 (SID20182516)



Average Marks per semester:

While it appears that the marks dropped significantly and have fluctuated throughout the years, it is important to note the scale of the average grade. The average has maintained fairly consistent around the 69% range, with it fluctuating between 69.45 and 69.70. This shows that students' average grades are continuing to be consistent regardless fo potential external factors.