Student Grades Data Analysis Report

The reason I personally chose this specific project is mainly because, as a master student, I wanted to see by myself how the final grade of a course can be affected during the semester by many factors that as a student, I have experienced and now I have the knowledge to apply.

Step by step process

1. Problem Understanding

The purpose of this project is to explore how student performance is influenced in a specific module by participation in Homework Assignments, Compulsory Activities, and Optional Activities. The project aims to discover patterns and relationships that explain students' success or failure in Exams and Repeated Exams, and to model performance outcomes using supervised and unsupervised learning techniques.

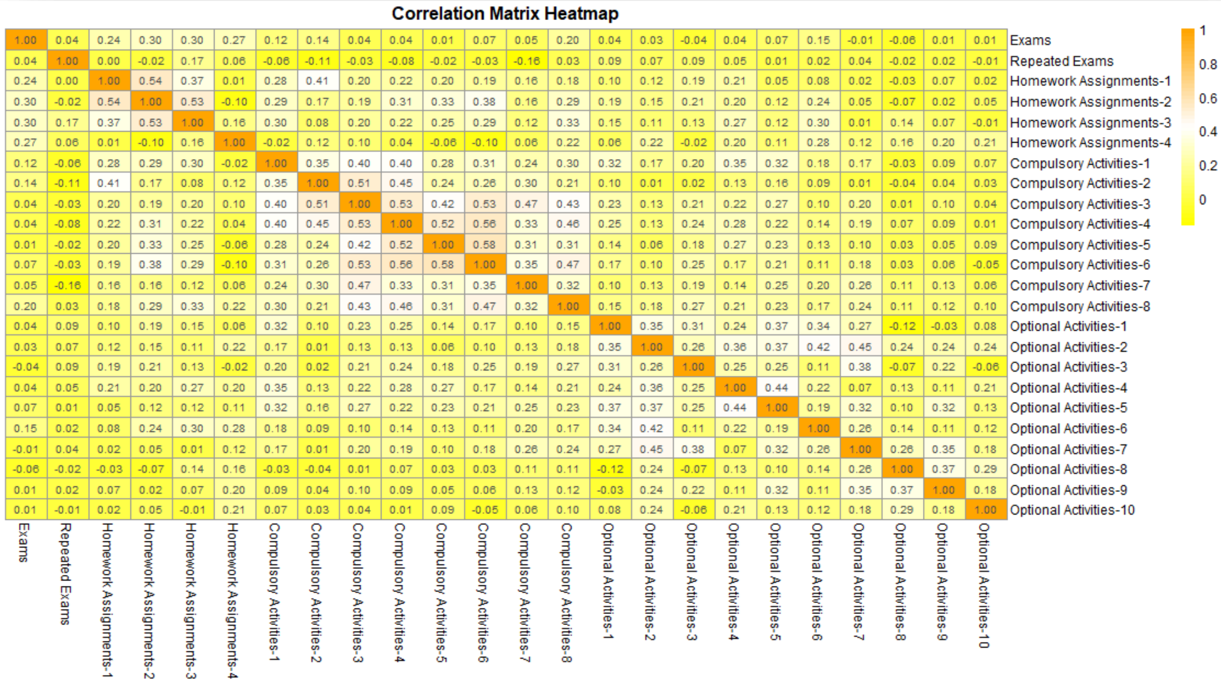
2. Data Understanding

The dataset includes performance data for 159 students. Key columns include participation in Homework Assignments, Compulsory Activities, Optional Activities, and their performance in Exams and Repeated Exams. **Early analysis revealed important distributions such as: 101 students took the first exam, 37 took the repeated exam, and 44 students did not participate in either**. Based now on simple data understanding we came up with some interesting early conclusions:

- ✓ Out of the only 47 students who completed all optional and compulsory activities,36 students succeeded on the final exams, which is a percentage of around 76%.
- ✓ Students who did not complete in any of the exams did not complete all Homework Assignments and Compulsory Activities.
- ✓ Interestingly, out of only 39 students who completed at least half of the optional activities, 38 succeeded on the final exams or repeated exams which is a percentage of around 97% (extremely high).

3. Data Preparation

Initial preparation involved:
- Removing unnecessary rows and columns
- Renaming columns
- Replacing values like -1 and 1 with NA where appropriate
- Creating binary indicator columns for full participation (binary.1) and partial participation (binary.2), with binary.1 indicating students who showed up for either exams or repeated exams and completed all Homework Assignments and Compulsory Activities and binary.2 which indicates with the value 1 students who completed at least half of the optional activities and with zero those that did not.
- Imputing NA values with column-wise means.
- Creating correlation heatmaps for visual insight. After creating the correlation heatmap, we end up with some interesting conclusions:



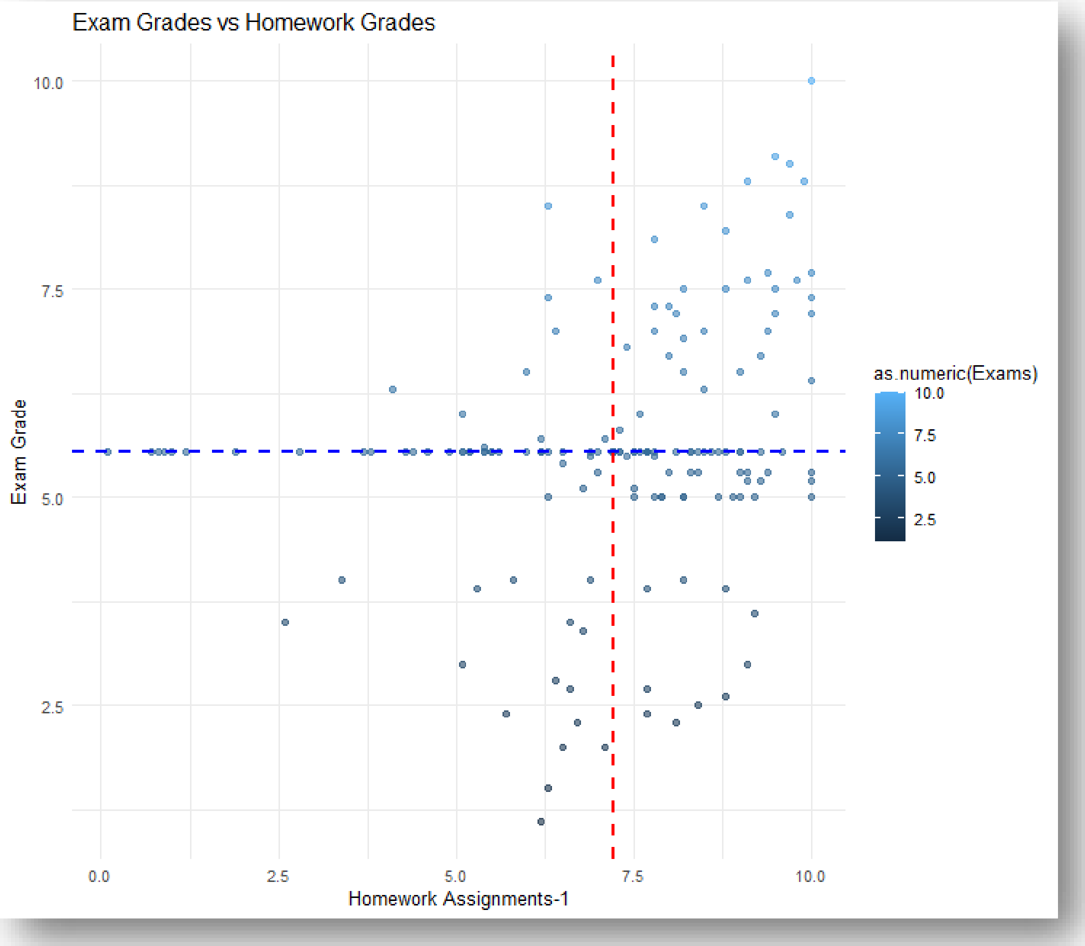
Correlation Matrix Heatmap

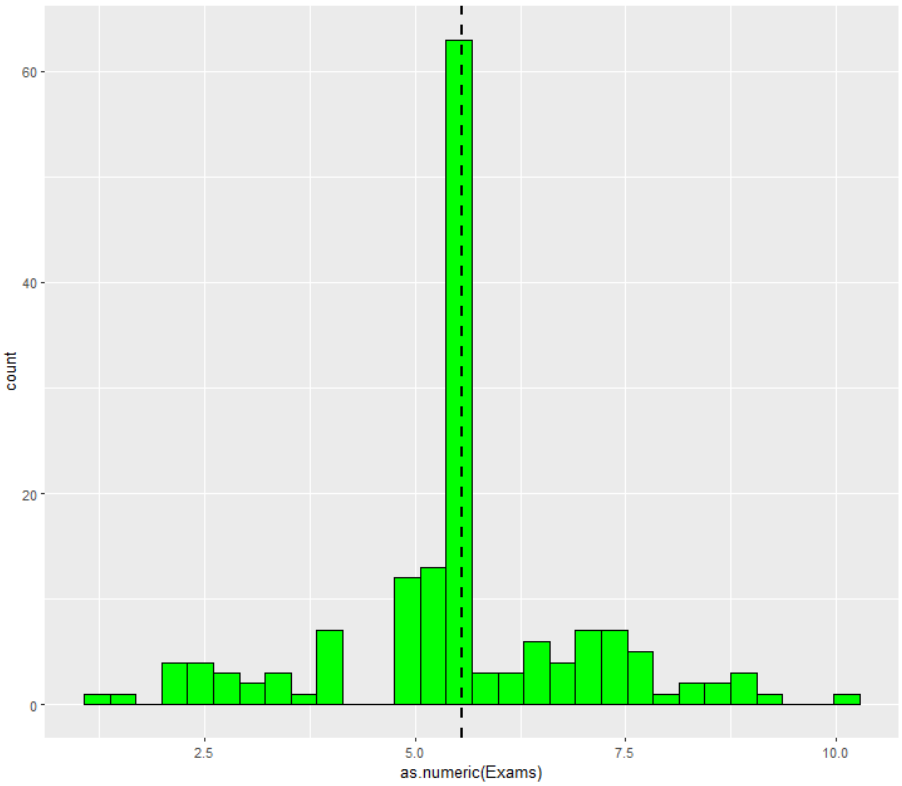
Exams	Repeated Exams	Homework Assignments-1	Homework Assignments-2	Homework Assignments-3	Homework Assignments-4	Compulsory Activities-1	Compulsory Activities-2	Compulsory Activities-3	Compulsory Activities-4	Compulsory Activities-5	Compulsory Activities-6	Compulsory Activities-7	Compulsory Activities-8	Optional Activities-1	Optional Activities-2	Optional Activities-3	Optional Activities-4	Optional Activities-5	Optional Activities-6	Optional Activities-7	Optional Activities-8	Optional Activities-9	Optional Activities-10	
1.00	0.04	0.24	0.30	0.30	0.27	0.12	0.14	0.04	0.04	0.01	0.07	0.05	0.20	0.04	0.03	-0.04	0.04	0.07	0.15	-0.01	-0.06	0.01	0.01	Exams
0.04	1.00	0.00	-0.02	0.17	0.06	-0.06	-0.11	-0.03	-0.08	-0.02	-0.03	-0.16	0.03	0.09	0.07	0.09	0.05	0.01	0.02	0.04	-0.02	0.02	-0.01	Repeated Exams
0.24	0.00	1.00	0.54	0.37	0.01	0.28	0.41	0.20	0.22	0.20	0.19	0.16	0.18	0.10	0.12	0.19	0.21	0.05	0.08	0.02	-0.03	0.07	0.02	Homework Assignments-1
0.30	-0.02	0.54	1.00	0.53	-0.10	0.29	0.17	0.19	0.31	0.33	0.38	0.16	0.29	0.19	0.15	0.21	0.20	0.12	0.24	0.05	-0.07	0.02	0.05	Homework Assignments-2
0.30	0.17	0.37	0.53	1.00	0.16	0.30	0.08	0.20	0.22	0.25	0.29	0.12	0.33	0.15	0.11	0.13	0.27	0.12	0.30	0.01	0.14	0.07	-0.01	Homework Assignments-3
0.27	0.06	0.01	-0.10	0.16	1.00	-0.02	0.12	0.10	0.04	-0.06	-0.10	0.06	0.22	0.06	0.22	-0.02	0.20	0.11	0.28	0.12	0.16	0.20	0.21	Homework Assignments-4
0.12	-0.06	0.28	0.29	0.30	-0.02	1.00	0.35	0.40	0.40	0.28	0.31	0.24	0.30	0.32	0.17	0.20	0.35	0.32	0.18	0.17	-0.03	0.09	0.07	Compulsory Activities-1
0.14	-0.11	0.41	0.17	0.08	0.12	0.35	1.00	0.51	0.45	0.24	0.26	0.30	0.21	0.10	0.01	0.02	0.13	0.16	0.09	0.01	-0.04	0.04	0.03	Compulsory Activities-2
0.04	-0.03	0.20	0.19	0.20	0.10	0.40	0.51	1.00	0.53	0.42	0.53	0.47	0.43	0.23	0.13	0.21	0.22	0.27	0.10	0.20	0.01	0.10	0.04	Compulsory Activities-3
0.04	-0.08	0.22	0.31	0.22	0.04	0.40	0.45	0.53	1.00	0.52	0.56	0.33	0.46	0.25	0.13	0.24	0.28	0.22	0.14	0.19	0.07	0.09	0.01	Compulsory Activities-4
0.01	-0.02	0.20	0.33	0.25	-0.06	0.28	0.24	0.42	0.52	1.00	0.58	0.31	0.31	0.14	0.06	0.18	0.27	0.23	0.13	0.10	0.03	0.05	0.09	Compulsory Activities-5
0.07	-0.03	0.19	0.38	0.29	-0.10	0.31	0.26	0.53	0.56	0.58	1.00	0.35	0.47	0.17	0.10	0.25	0.17	0.21	0.11	0.18	0.03	0.06	-0.05	Compulsory Activities-6
0.05	-0.16	0.16	0.16	0.12	0.06	0.24	0.30	0.47	0.33	0.31	0.35	1.00	0.32	0.10	0.13	0.19	0.14	0.25	0.20	0.26	0.11	0.13	0.06	Compulsory Activities-7
0.20	0.03	0.18	0.29	0.33	0.22	0.30	0.21	0.43	0.46	0.31	0.47	0.32	1.00	0.15	0.18	0.27	0.21	0.23	0.17	0.24	0.11	0.12	0.10	Compulsory Activities-8
0.04	0.09	0.10	0.19	0.15	0.06	0.32	0.10	0.23	0.25	0.14	0.17	0.10	0.15	1.00	0.35	0.31	0.24	0.37	0.34	0.27	-0.12	-0.03	0.08	Optional Activities-1
0.03	0.07	0.12	0.15	0.11	0.22	0.17	0.01	0.13	0.13	0.06	0.10	0.13	0.18	0.35	1.00	0.26	0.36	0.37	0.42	0.45	0.24	0.24	0.24	Optional Activities-2
-0.04	0.09	0.19	0.21	0.13	-0.02	0.20	0.02	0.21	0.24	0.18	0.25	0.19	0.27	0.31	0.26	1.00	0.25	0.25	0.11	0.38	-0.07	0.22	-0.06	Optional Activities-3
0.04	0.05	0.21	0.20	0.27	0.20	0.35	0.13	0.22	0.28	0.27	0.17	0.14	0.21	0.24	0.36	0.25	1.00	0.44	0.22	0.07	0.13	0.11	0.21	Optional Activities-4
0.07	0.01	0.05	0.12	0.12	0.11	0.32	0.16	0.27	0.22	0.23	0.21	0.25	0.23	0.37	0.37	0.25	0.44	1.00	0.19	0.32	0.10	0.32	0.13	Optional Activities-5
0.15	0.02	0.08	0.24	0.30	0.28	0.18	0.09	0.10	0.14	0.13	0.11	0.20	0.17	0.34	0.42	0.11	0.22	0.19	1.00	0.26	0.14	0.11	0.12	Optional Activities-6
-0.01	0.04	0.02	0.05	0.01	0.12	0.17	0.01	0.20	0.19	0.10	0.18	0.26	0.24	0.27	0.45	0.38	0.07	0.32	0.26	1.00	0.26	0.35	0.18	Optional Activities-7
-0.06	-0.02	-0.03	-0.07	0.14	0.16	-0.03	-0.04	0.01	0.07	0.03	0.03	0.11	0.11	-0.12	0.24	-0.07	0.13	0.10	0.14	0.26	1.00	0.37	0.29	Optional Activities-8
0.01	0.02	0.07	0.02	0.07	0.20	0.09	0.04	0.10	0.09	0.05	0.06	0.13	0.12	-0.03	0.24	0.22	0.11	0.32	0.11	0.35	0.37	1.00	0.18	Optional Activities-9
0.01	-0.01	0.02	0.05	-0.01	0.21	0.07	0.03	0.04	0.01	0.09	-0.05	0.06	0.10	0.08	0.24	-0.06	0.21	0.13	0.12	0.18	0.29	0.18	1.00	Optional Activities-10

- **No significant influence of any of Optional, Compulsory and Homework Assignments to the Repeated Exams.**
- **Only Homework Assignments have an impact on Exams, yet not significant, with the highest being 0.30.**
- **Some significant correlations between the Homework Assignments (between them) and the compulsory Activities (between them).**

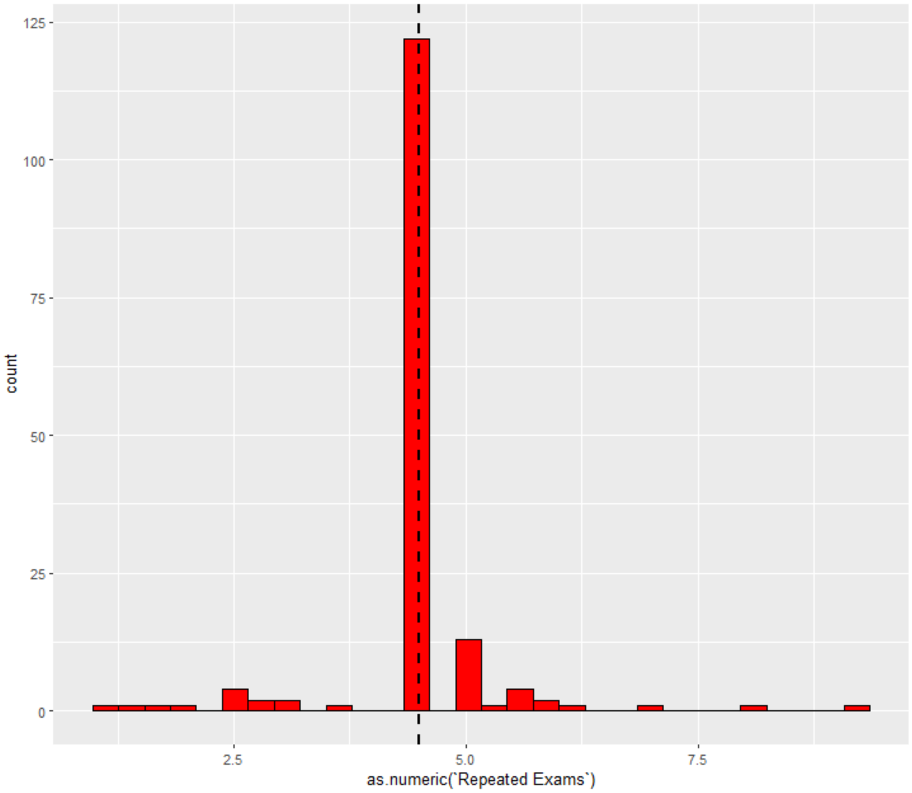
-After, some scatterplots and histograms where also created for visual insight based on which some conclusions can be noticed:

- **We can see that in the Exams the majority of grads is gathered around 5.0 and 7.5**
- **In Repeated Exams the majority of grades is gathered in 5.**
- **Based on the scatterplots, we can see some correlation between Exams and Homework Assignments.**





60
40
20
0
count
2.5
5.0
7.5
10.0
as.numeric(Exams)

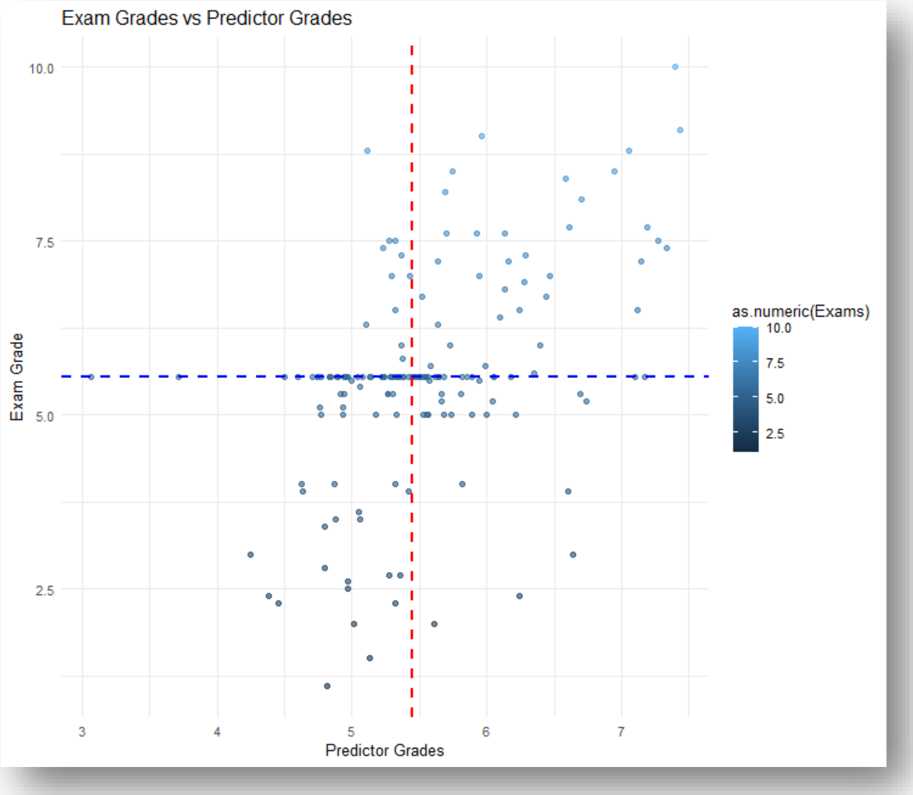


125
100
75
50
25
0
count
2.5
5.0
7.5
as.numeric(`Repeated Exams`)

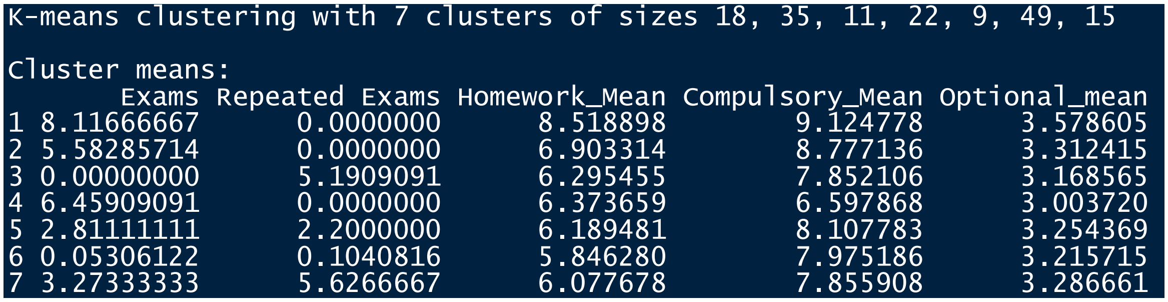
5. Finding Explanations

Several models were applied:

- KNN: Achieved 83.3% accuracy on a subset of 48 students predicting pass/fail status for Exams
- Linear Regression: Low R^2 (~22%), significant variables were Homework Assignment 2 and 4. The image below shows a scatterplot between the Actual Exam values and the predicted ones.



- Logistic Regression: Predictive accuracy up to 93%, but poor performance on 'fail' cases
- K-Means Clustering (k=7): Grouped students by performance categories. Cluster 1 contained top performers; Cluster 6 had students who didn't show up; Cluster 5 grouped students with poor outcomes. Cluster 2 and 4 separated the good students properly (cluster 4 good students and cluster 2 the semi-good students). At the same time, it groups at cluster 3 all the students who managed to succeed during the repeated exams without having taken place on the exams.



```
K-means clustering with 7 clusters of sizes 18, 35, 11, 22, 9, 49, 15

Cluster means:
       Exams Repeated Exams Homework_Mean Compulsory_Mean Optional_mean
1 8.11666667      0.0000000      8.518898        9.124778      3.578605
2 5.58285714      0.0000000      6.903314        8.777136      3.312415
3 0.00000000      5.1909091      6.295455        7.852106      3.168565
4 6.45909091      0.0000000      6.373659        6.597868      3.003720
5 2.81111111      2.2000000      6.189481        8.107783      3.254369
6 0.05306122      0.1040816      5.846280        7.975186      3.215715
7 3.27333333      5.6266667      6.077678        7.855908      3.286661
```

6. Conclusions and Insights

Key takeaways include:

- Participation in Optional Activities is a strong predictor of exam success.
- 44 students skipped both exams, but many had decent participation in activities, indicating a possible motivational or structural issue.
- Linear models do not explain much variance; logistic models perform better but still fail in edge cases.
- K-Means clustering effectively grouped students with similar profiles, including high-achievers and at-risk students.

Recommendations include: increasing the impact of activity grades, simplifying assignment load, and collecting qualitative feedback from students about the course.

7. Tools Used

The analysis was performed using the following tools and libraries:

- R programming language
- ggplot2, dplyr, and caret for data manipulation and modeling
- base R functions for statistical processing
- heatmap() and kmeans() for visualizations and clustering

-Logistic Regression models and Linear Regression Models

References

1. R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/
2. Wickham, H. (2023). *ggplot2: Elegant Graphics for Data Analysis*. Springer. https://ggplot2.tidyverse.org/

3. Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org/
4. Kuhn, M. (2023). *caret: Classification and Regression Training*. https://topepo.github.io/caret/
5. R Documentation. (2024). *Base R Statistical Functions*. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
6. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, 2825–2830. https://scikit-learn.org/stable/modules/neighbors.html
7. Wikipedia contributors. (2024). *Linear regression*. Wikipedia. https://en.wikipedia.org/wiki/Linear_regression
8. Wikipedia contributors. (2024). *Logistic regression*. Wikipedia. https://en.wikipedia.org/wiki/Logistic_regression
9. Wikipedia contributors. (2024). *K-means clustering*. Wikipedia. https://en.wikipedia.org/wiki/K-means_clustering
10. R Graph Gallery. (2024). *Correlation heatmaps in R*. https://r-graph-gallery.com/79-correlation-heatmap.html