Tecnológico de Monterrey

Campus Ciudad de México

Activity 4: Heatmaps and boxplots

Ximena Silva Bárcena

A01785518

Rodrigo Martínez Vallejo

A00573055

Prof. Sergio Ruiz Loza

Mastering Analytics

Grupo **201**

Wednesday, May 7th, 2025

Ximena Silva Bárcena
Are there any variables that do not provide information?
Variables like parental_education_level and internet_quality show very low correlation with the exam_score and other key performance indicators. This suggests they might not significantly influence the overall academic performance of student
If you had to eliminate variables, which ones would you remove and why?
Parental_education_level and internet_quality might also be candidates for removal if their correlation with academic performance is weak or if they introduce unnecessary complexity without significantly contributing to the model's predictive power.
Are there any variables with unusual data?
The mental_health_rating variable shows a wide range with extreme low and high values, which might indicate potential outliers or inconsistent data entries. The attendance_percentage also appears to have values close to 100, suggesting it might be skewed or have limited variance.
If you compare the variables, are they all in similar ranges?
The variables have different scales. For example, sleep_hours ranges from 3.2 to 10, while attendance_percentage ranges from 0 to 100. This wide variation in scales can distort distance-based algorithms like K-Means, leading to misleading clustering results if not properly normalized.
Do you think this affects the data analysis? Can you find any similar groups? What are these groups?
Using unscaled data can lead to misleading results in cluster analysis because features with larger numerical ranges dominate the distance calculations. For example, early cluster tests might incorrectly separate students based on attendance rates or sleep hours alone, ignoring other critical factors like study

habits or mental health. To avoid this, it's crucial to normalize the data, revealing more accurate groupings, like diligent students with high exam scores and balanced lifestyles, versus distracted students with high entertainment consumption and lower academic outcomes.

Rodrigo Martínez Vallejo
Are there any variables that do not provide information?
Parental Education Level and Internet Quality both show very weak correlations with exam scores, making them poor predictors of academic success. Similarly, Extracurricular Participation has almost no correlation with exam results, suggesting it doesn't contribute valuable information to the analysis.
If you had to eliminate variables, which ones would you remove and why?
I would consider removing Parental Education Level, Internet Quality, and Extracurricular Participation because their weak correlations with exam scores indicate that they add little to no predictive power. Removing these variables can reduce model complexity and improve accuracy.
Are there any variables with unusual data?
Netflix Hours and Social Media Hours have extreme maximum values that likely indicate outliers or data entry mistakes, as it's unlikely that students spend this much time daily without severe academic impact.
If you compare the variables, are they all in similar ranges?
No, the ranges vary significantly. For example, attendance_percentage ranges from 0 to 100, while exercise_frequency is capped at 7. This imbalance can distort machine learning algorithms that rely on distance calculations, potentially skewing clustering results.

Do you think this affects the data analysis? Can you find any similar groups? What are these groups?
The differences in scales can affect data analysis by giving disproportionate weight to certain features. For instance, attendance percentages might overpower other metrics if not properly scaled. Preliminary clustering attempts suggest the existence of at least two main groups: students who study consistently with strong mental health and high scores, and another group characterized by excessive social media use, poor exam performance, and lower attendance.