Group 14 - Overview of Updates Since Previous Meeting (June 27th)

Notebook 1. Data Cleaning

- **Column: Rating**
  - **Previous Approach:** Halved ratings due to values exceeding 5 and rounded ratings to one decimal place.
  - **Updated Approach:** Dropped values above 5 (only 3 entries) and added a new column “rating_round” for visualization; the original “rating” column is used for machine learning purposes.
- **Column: Number of Employees**
  - **Previous Approach:** Retained stores with 0 employees, assuming they were kiosks or vending machines.
  - **Updated Approach:** Updated 0 employee entries to reflect the average number for stores of 400 m², as kiosks/vending machines are unlikely to occupy such space.
- **Column: Parking**
  - **Previous Approach:** Assumed parking value error: “10” == 1 and “-1” == 0.
  - **Updated Approach:** Dropped values “10” and “-1” as they only had 2 entries.

Notebook 2. Feature Engineering

- **Column: Store Sub-Category**
  - **Previous Approach:** Used manual classification and Regex for information extraction.
  - **Updated Approach:** Combined manual rule-based classification with the T5 model for ambiguous cases.
    - Note: Attempted to use the GPT-3.5-turbo for classification, but it was too time-consuming thus not implemented. Currently still working on implementing RAG
- **Add Column: Sales Rep Population**
  - **Purpose:** Calculate the percentage of the population managed by each sales representative based on the number of unique stores they oversee.

Notebook 3. Exploratory Data Analysis (EDA)

- The primary information remains consistent. The presentation is now organized into six sections for better structure. More details are available in the notebook.

Notebook 4. Regression Model

- **Previous Approach:** Utilized multiple regression models (Linear Regression, KNN, CatBoost, XGBoost) and grid search.
- **Updated Approach:** Focused on linear regression with feature selection via AIC/BIC.