**Final Project Summary — Clinic Appointments No-Show Analysis**

The purpose of this project was to analyze a dataset of clinic appointments to better understand the reasons behind patient “no-shows.” By carefully cleaning, preparing, and exploring the data, the analysis aimed to uncover patterns that influence attendance and suggest strategies that clinics can use to reduce missed appointments.

The dataset contained approximately 5,400 rows, each representing one appointment over a ninety-day period. Key features included patient age, gender, distance from the clinic, prior no-shows, appointment lead time, SMS reminders, whether the appointment was scheduled on a weekend, and whether it occurred on a rainy day. The target variable was no\_show, coded as 1 for patients who did not attend and 0 for those who did.

The first stage, Milestone 2, involved exploratory data analysis. This step used pandas functions to examine the structure and characteristics of the data, such as .info(), .describe(), and .head(). The results confirmed that the dataset was well structured, free of missing values and duplicates, and suitable for further preparation. This initial exploration also provided an overview of distributions and revealed potential inconsistencies, such as variation in gender labels, which would need to be standardized.

In Milestone 3 the focus shifted to data cleaning. The date column was converted to a datetime type so that time-based patterns could be analyzed accurately. Gender labels were standardized into the categories Male, Female, and Other to eliminate inconsistencies. Binary variables, including weekend, SMS reminder, rain, and no\_show, were coerced into integers with only 0 and 1 values. Numeric columns such as age, distance\_km, lead\_time\_days, and prior\_no\_shows were converted to numeric types with invalid entries filled as zero. Outliers were addressed through IQR-based winsorizing, which capped extreme values while retaining all rows. Logical validity checks ensured that ages were realistic, distances and counts were nonnegative, and binary variables only contained allowable values. At the end of this stage the cleaned dataset was verified and ready for analysis.

Milestone 4 carried out the analysis and visualization. The overall no-show rate was calculated, showing that roughly X percent of all appointments were missed. Further groupings highlighted the main drivers of no-shows. Patients who received SMS reminders were significantly less likely to miss their appointments compared to those who did not receive reminders. Appointments with longer lead times, particularly those scheduled more than a week in advance, had a noticeably higher no-show rate. Gender showed only minor differences, indicating it was not a strong factor in attendance. Bar charts were used to display these patterns clearly, reinforcing the observed relationships between these variables and the target outcome.

The findings of this project suggest two key areas where clinics can take action. Consistently sending SMS reminders helps lower the chance of missed appointments, while keeping lead times shorter reduces the likelihood of patients forgetting or cancelling. Gender does not appear to meaningfully affect attendance, which highlights that operational changes rather than demographic factors are the most effective levers.

Overall, this project demonstrates the value of structured data analysis in improving healthcare operations. By following the complete process of exploration, cleaning, and analysis, it was possible to generate insights that can guide scheduling policies and communication strategies, ultimately reducing no-show rates and improving both efficiency and patient care.