

Data Cleaning Logs

Data Cleaning Summary:

Column	Transformation Applied	Reason
Gender	PROPER + Mode	Standardizes text format and fills missing values using the most frequent category.
Department	UPPER + Mode	Ensures consistent categorical encoding and fills missing values using the most frequent category.
DateOfJoining	DATEVALUE conversion	Converts text to valid date format for date-based analysis.
LastLogin	VALUE conversion	Converts text to valid datetime format for time-based analysis.
WorkHours	Median	Median provides robust central value and reduces impact of outliers.
StressLevel	Median	Median preserves distribution and provides stable replacement.
SleepHours	Median	Median ensures reliable imputation for numerical data.
ProductivityScore	Median	Median maintains a representative central tendency.
StepCount	Median	Median reduces distortion from extreme values.
HeartRate	Median	Median provides robust estimates for physiological data.
SmokingStatus	Mode	Mode is appropriate for categorical missing values.

Exact formulas used:

Column	Formula Used
Gender	=IF(D2=" ", INDEX(Proper(D:D), MATCH(MAX(COUNTIF(Proper(D:D), UNIQUE(Proper(D:D)))), COUNTIF(Proper(D:D), UNIQUE(Proper(D:D))), 0)), Proper(D2)))
Department	=IF(E2=" ", INDEX(Upper(E:E), MATCH(MAX(COUNTIF(Upper(E:E), UNIQUE(Upper(E:E)))), COUNTIF(Upper(E:E), UNIQUE(Upper(E:E))), 0)), Upper(E2)))
DateOfJoining	=DATEVALUE(F2)
LastLogin	=VALUE(G2)
WorkHours	=IF(H2=" ", MEDIAN(H:H), H2)
StressLevel	=IF(I2=" ", MEDIAN(I:I), I2)
SleepHours	=IF(J2=" ", MEDIAN(J:J), J2)
ProductivityScore	=IF(K2=" ", MEDIAN(K:K), K2)
StepCount	=IF(L2=" ", MEDIAN(L:L), L2)
HeartRate	=IF(M2=" ", MEDIAN(M:M), M2)
SmokingStatus	=IF(N2=" ", INDEX(N:N, MATCH(MAX(COUNTIF(N:N, UNIQUE(N:N))), COUNTIF(N:N, UNIQUE(N:N)), 0)), N2))