Complete Description of the `/reports` Folder

File Name	Type	Description	Main Purpose	For ML Engineer	For Final Report
kaggle_clean_drop_outliers.csv	CSV	Kaggle Sleep Health dataset after removing outliers using the IQR method.	Cleaned training data for testing model robustness.	✅	
kaggle_clean_winsorized.csv	CSV	Dataset after Winsorization (trimming extreme values).	Alternative to outlier removal, for comparing data-cleaning strategies.	✅	
kaggle_cleaned_snapshot.csv	CSV	Final standardized cleaned dataset with consistent column names, types, and formats.	Main cleaned dataset for modeling and reporting.	✅	✅
kaggle_iqr_outliers.csv	CSV	Records of all samples identified as outliers using the IQR method.	Shows distribution of anomalies before cleaning.		✅
kaggle_missingness.csv	CSV	Summary table showing missing-value percentages per variable.	Used in data quality and completeness analysis.		✅
kaggle_numeric_summary.csv	CSV	Statistical summary (mean, std, IQR, min, max) of all numerical variables.	Describes data distribution in the report.		✅
outcome_week9.txt	TXT	Week 09 progress summary (Data Cleaning + Preprocessing).	Team milestone documentation or report appendix.		✅
preprocess_feature_names.csv	CSV	Complete mapping of feature names after ColumnTransformer (including One-Hot encoded columns).	Helps model interpretation and feature-importance debugging.	✅	
preprocessor.joblib	Binary	Trained preprocessing pipeline (includes scaler, encoder, imputer, etc.).	Enables ML engineer to load and apply identical transformations.	✅	
X_train_proc.parquet	Parquet	Preprocessed training feature matrix (standardized/encoded).	Input for model training.	✅	
X_test_proc.parquet	Parquet	Preprocessed testing feature matrix.	Input for model validation.	✅	
y_train.csv	CSV	Training labels (sleep quality).	Target variable for training.	✅	
y_test.csv	CSV	Testing labels (sleep quality).	Target variable for validation.	✅	

y_test.csv	CSV	Testing labels (sleep quality).	Target variable for validation.	✅	
norm_stats_minmax.csv	CSV	Normalization parameters (min, max) used by MinMaxScaler.	Records scaling parameters for transparency.		✅
norm_stats_robust.csv	CSV	Median/IQR parameters used by RobustScaler.	For comparing scaling effects between methods.		✅
norm_stats_standard.csv	CSV	Mean and standard deviation used by StandardScaler.	Supports explanation of normalization scheme in the report.		✅
data_dictionary.csv	CSV	Data dictionary with column names, meanings, data types, and units.	Clarifies variable definitions for modeling and documentation.	✅	✅
data_dictionary.md	Markdown	Same as above, but in descriptive text format with data source notes.	For inclusion in reports and project documentation.	✅	✅
README_preprocessing.md	Markdown	Complete documentation of the data-cleaning and preprocessing workflow.	Handoff note ensuring full reproducibility.	✅	✅

📦 Summary Table by Category

Category	Files to Provide to ML Engineer	Description
✅ **Core Modeling Files (Required)**	`preprocessor.joblib`, `X_train_proc.parquet`, `y_train.csv`, `X_test_proc.parquet`, `y_test.csv`, `preprocess_feature_names.csv`	Can be loaded directly for modeling without re-cleaning.
🟦 **Supporting Files (Recommended)**	`data_dictionary.csv`, `data_dictionary.md`, `README_preprocessing.md`, `kaggle_cleaned_snapshot.csv`, `kaggle_clean_drop_outliers.csv`, `kaggle_clean_winsorized.csv`	Provide context for data background and cleaning alternatives.
📊 **Report-Only Files (For Final Report)**	`kaggle_missingness.csv`, `kaggle_numeric_summary.csv`, `kaggle_iqr_outliers.csv`, `norm_stats_standard.csv`, `norm_stats_minmax.csv`, `norm_stats_robust.csv`, `outcome_week9.txt`	Used for writing the Data Analysis and Methodology sections.

✅ Summary of Deliverables Completed

Project Stage	Status	Outputs
Data Collection (Sleep-EDF + Kaggle)	✅ Completed	Raw CSV / EDF files (not uploaded)
Data Cleaning & Standardization	✅ Completed	Various `kaggle_clean*.csv`, `norm_stats_*.csv`
Missing Value & Outlier Handling	✅ Completed	`kaggle_missingness.csv`, `kaggle_iqr_outliers.csv`
Preprocessing Pipeline Construction	✅ Completed	`preprocessor.joblib`, `README_preprocessing.md`
Train–Test Data Export	✅ Completed	`X_train_proc.parquet`, `X_test_proc.parquet`, `y_train.csv`, `y_test.csv`
Data Documentation	✅ Completed	`data_dictionary.csv`, `data_dictionary.md`, `outcome_week9.txt`