FIRST RECIPE
1st Prompt
Accuracy 8/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit check on target variable distribution like `value_counts()` or `countplot()`
Sampling type	Random	`train_test_split(..., random_state=42)` used without `stratify`
Outliers removal	Yes	Z-score plots visualized for each numeric column with threshold ±3, no rows dropped but flagged visually
Check for duplicates	Yes	Used `df.duplicated().value_counts()` to count duplicates
Imputation of missing values	none	All columns showed 0 missing values; no imputation was applied
Drop columns	Yes	Dropped `No`, `y_house_price_of_unit_area`, and latitude/longitude during processing
Encoding	Label Encoder	Applied `LabelEncoder()` on `X1 transaction date` and cluster labels
Create new columns	Yes	Created cluster, cluster_density, and other engineered features
Feature selection	Yes	Selected subset of columns for modeling: e.g., MRT distance, convenience stores, cluster, and density
Data scaling/standardisation	Yes	Used `StandardScaler()` on selected features
Hyperparameter tuning	Yes	Used `GridSearchCV` and `RandomizedSearchCV` for multiple models

2nd Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details

Check for balanced data	No	No explicit check of target variable distribution via value_counts or plots
Sampling type	Random	`train_test_split` with random state used multiple times
Outliers removal	Yes	Z-score plots visualised outliers (±3), though not removed programmatically
Check for duplicates	Yes	`df.duplicated().value_counts()` used
Imputation of missing values	None	Missing values were analyzed but not imputed; no `.fillna()` or similar used
Drop columns	Yes	`'No'` column and target column dropped completely without replacement
Encoding	Label Encoder	Label encoding used for transaction date and cluster labels
Create new columns	Yes	New columns created from clustering and feature engineering, e.g., `cluster`, `*_density`
Feature selection	Yes	Features dropped before training; correlation heatmaps used
Data scaling/standardisation	Yes	`StandardScaler` applied before model fitting
Hyperparameter tuning	Yes	Both `GridSearchCV` and `RandomizedSearchCV` used for tuning multiple models

3rd Prompt
Accuracy 8/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit check for distribution of target or labels using `value_counts` or visualizations.
Sampling type	Random	Used `train_test_split(..., random_state=42)` without stratification or oversampling.

Outlier removal	Yes	Outliers visualised using Z-score plots, though not explicitly removed. Based on decision rule, visualisation counts as indication: Yes.
Check for duplicates	Yes	`df.duplicated().value_counts()` is used to check for duplicates.
Imputation of missing values	No	Missing values are analysed using `df.isna().sum()`, but no imputation is performed.
Drop columns	Yes	Dropped columns like `No` and `y_house_price_of_unit_area`.
Encoding	Label Encoder	Used `LabelEncoder` for `'X1 transaction date'` and clustering.
Create new columns	Yes	New columns like `'cluster'`, `'x3_distance_to_the_nearest_mrt_station_density'`, and polynomial features were created.
Feature selection	Yes	Features were selected manually for `X` (e.g., dropping latitude/longitude and keeping 4 predictors only).
Standardization	Yes	Used `StandardScaler` for feature scaling.
Hyperparameter tuning	Yes	Used `GridSearchCV` and `RandomizedSearchCV` for multiple models including XGB, SVR, Ridge, Lasso, CatBoost.

4th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No countplot, value_counts, or distribution check for the target variable
Sampling type	Random	`train_test_split(..., random_state=42)` without `stratify` used multiple times
Outliers removal	No	Z-score plotted but no rows removed or filtered based on threshold

Check for duplicates	Yes	`df.duplicated().value_counts()` used to count duplicates
Imputation of missing values	none	No imputation methods (`fillna`, dropna, SimpleImputer, etc.) applied
Drop columns	Yes	Columns dropped (`'No'`, target variable)
Encoding	Label Encoder	`LabelEncoder` used on `X1 transaction date` and `cluster` columns
Create new columns	Yes	`cluster`, `density`, and Dash input-based predictions use external info, not derived only
Feature selection	Yes	Correlation and domain insights used to retain only 4 features for model input
Data scaling/standardisation	Yes	`StandardScaler` applied before modeling
Hyperparameter tuning	Yes	`GridSearchCV` and `RandomizedSearchCV` used for multiple models

5th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value_counts/countplot/hist on target variable (`y_house_price_of_unit_area`) observed.
Sampling type	Random	`train_test_split(..., stratify=...)` not used; standard random split is used multiple times.
Outliers removal	Yes	Z-score plots with threshold ±3 used to visualize and identify outliers.
Check for duplicates	Yes	`df.duplicated().value_counts()` used to assess duplicates.
Imputation of missing values	none	No explicit handling or imputation (ignore/drop/replace) of missing values observed. All features had 0 missing.

Drop columns	Yes	`'No'` column and target column dropped completely without replacement
Encoding	Label Encoder	`LabelEncoder` used for encoding `X1 transaction date` and clustering label.
Create new columns	No	All derived columns (log transform, clustering, density) are based on existing columns.
Feature selection	Yes	Only selected features (`distance`, `stores`, `cluster`, `density`) used for model training; rest dropped after correlation/EDA insights.
Data scaling/standardisation	Yes	`StandardScaler` applied before polynomial feature generation.
Hyperparameter tuning	Yes	`GridSearchCV` and `RandomizedSearchCV` used on multiple models (XGBoost, SVR, Ridge, etc.) with defined parameter grids.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	Regression task; no explicit class-balance check on a target label (only numeric histograms/EDA).
Sampling type	Random	`train_test_split(X_poly, y, test_size=0.3, random_state=42)` without `stratify`.
Outliers removal	No	Only z-score plots created; no filtering/masking applied to remove rows.
Check for duplicates	Yes	Duplicate inspection via `df.duplicated().value_counts()`; no drops performed.
Imputation of missing values	none	Missingness profiled with `df.isna().sum()`; no `fillna`/imputer or row/column drops for NaNs.
Drop columns	Yes	Pre-EDA removal of identifier column No with `df.drop(columns=['No'])` (not reused elsewhere).

Encoding	Label Encoder	`LabelEncoder` applied to `'X1 transaction date'`; `LabelEncoder` also used on KMeans `'cluster'` labels.
Create new columns	No	New fields (`cluster`, density, polynomial features) are derived from existing data (KMeans/groupby/poly), which does **not** count as “new” under the rubric.
Feature selection	Yes	Post-EDA **manual subset** chosen for modeling: `['x3_distance_to_the_nearest_mrt_station','x4_number_of_convenience_stores','cluster','x3_distance_to_the_nearest_mrt_station_density']`.
Data scaling/standardisation	Yes	`StandardScaler` fit and applied to features before modeling.
Hyperparameter tuning	Yes	Systematic search via `GridSearchCV` (cv=5) over defined parameter grids for multiple models.

SECOND RECIPE
1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit check (e.g. `value_counts()` or histogram) on the distribution of the target variable
Sampling type	Random	Used `train_test_split(..., random_state=42)` without `stratify`
Outliers removal	Yes	Applied IQR-based clipping on `X3 distance to MRT station` and `X6 longitude`
Check for duplicates	No	No check using `duplicated()` or similar method
Imputation of missing values	none	Checked for nulls with `isnull().sum()`, but didn’t apply any imputation
Drop columns	Yes	Dropped column `'No'` explicitly

Encoding	none	No encoding performed; no categorical columns were present or transformed
Create new columns	No	No derived or added features
Feature selection	Yes	Used all columns except dropped target; implicitly selected features through `drop(columns=...)`
Data scaling/standardisation	Yes	Applied `StandardScaler()` on features
Hyperparameter tuning	No	Direct instantiation of models; no grid/random search or tuning applied

2nd Prompt

Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check like `value_counts()` or distribution plot for target variable
Sampling type	Random	`train_test_split` with `random_state=42` used
Outliers removal	Yes	IQR method used to clip outliers in MRT station distance and longitude
Check for duplicates	No	No use of `df.duplicated()` or `.drop_duplicates()`
Imputation of missing values	None	Checked for missing values, but did not fill, drop, or impute
Drop columns	Yes	Column `'No'` dropped without replacement
Encoding	None	No categorical columns encoded
Create new columns	No	No new columns created; transformations stayed within original columns
Feature selection	No	No dropping of features based on importance, correlation, etc.
Data scaling/standardisation	Yes	`StandardScaler` used for `X_train` and `X_test`

Hyperparameter tuning	No	Models trained with fixed parameters; no grid/random search

3rd Prompt

Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No use of `value_counts`, histograms, or visual checks for target balance.
Sampling type	Random	Used `train_test_split(..., random_state=42)` without stratification or resampling.
Outlier removal	Yes	Handled using IQR method and capping for `'X3 distance to the nearest MRT station'` and `'X6 longitude'`.
Check for duplicates	No	No explicit check using `df.duplicated()` or similar.
Imputation of missing values	No	Checked `df.isnull().sum()`, but no imputation done.
Drop columns	Yes	Dropped `'No'` and `'Y house price of unit area'` during feature-target split.
Encoding	None	No categorical columns encoded. All features were numeric.
Create new columns	No	No creation of new columns observed.
Feature selection	No	All features retained (except target column). No selection based on correlation, importance, or filter methods.
Standardization	Yes	Used `StandardScaler` on training and test features.
Hyperparameter tuning	No	Models like RandomForestRegressor are used with manual parameters, but no GridSearch or tuning framework applied.

4th Prompt

Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value_counts or distribution plot for target variable
Sampling type	Random	`train_test_split(..., random_state=42)` used without stratification
Outliers removal	Yes	IQR-based capping applied to X3 `distance...` and X6 `longitude`
Check for duplicates	No	No check using `duplicated()` or similar
Imputation of missing values	none	`df.isnull().sum()` called, but no imputation or row/column drops performed
Drop columns	Yes	Column `'No'` dropped, which was an ID column
Encoding	none	No categorical columns encoded (none existed or needed)
Create new columns	No	No columns added beyond original ones
Feature selection	No	All features used except ID column
Data scaling/standardisation	Yes	`StandardScaler` applied to both train and test features
Hyperparameter tuning	No	Models manually specified with parameters (e.g., `RandomForestRegressor(...)`)

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check on the distribution of the target variable `Y house price of unit area` using value_counts or plots.
Sampling type	Random	`train_test_split` used without `stratify` → random sampling.

Outliers removal	Yes	IQR method applied on `X3 distance to the nearest MRT station` and `X6 longitude`; replaced outliers with threshold values.
Check for duplicates	No	No `duplicated()` check or similar method used.
Imputation of missing values	none	No imputation or handling required as `.isnull().sum()` showed zero missing values.
Drop columns	Yes	Column `'No'` was dropped but it's an ID column
Encoding	none	No categorical encoding was performed (all columns are numerical).
Create new columns	No	No new feature engineering or column creation beyond outlier adjustment.
Feature selection	No	All columns (except ID and target) were retained and used for modeling; no correlation-based or model-based selection performed.
Data scaling/standardisation	Yes	`StandardScaler` applied on features using `fit_transform` and `transform`.
Hyperparameter tuning	No	Models used with fixed/default hyperparameters; no GridSearchCV, RandomizedSearchCV, or manual param search observed.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit distribution check for the target (no `value_counts`/countplot/hist on `Y house price of unit area`).
Sampling type	Random	`train_test_split(..., test_size=0.33, random_state=42)` without `stratify`.
Outliers removal	Yes	Outliers are **winsorized/capped** via IQR limits for `X3 distance to the nearest MRT station` and `X6 longitude`

Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()` found.
Imputation of missing values	none	Missingness inspected (`isnull().sum()`), but no `fillna`/imputer or row/column drops applied.
Drop columns	Yes	Pre-EDA drop of identifier column `No` via `df.drop(columns=['No'], inplace=True)` (not reused elsewhere).
Encoding	none	Dataset is numeric; no `LabelEncoder`, `OneHotEncoder`, or `get_dummies` used.
Create new columns	No	No new features created; transformations are in-place caps on existing columns.
Feature selection	No	No correlation/variance/model-importance-based pruning or post-EDA drops.
Data scaling/standardisation	Yes	`StandardScaler` fitted on `X_train` and applied to `X_test`.
Hyperparameter tuning	No	Models trained with fixed/default params; no `GridSearchCV`/`RandomizedSearchCV`/Optuna used.

THIRD RECIPE
1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check like `value_counts()` or visualisation for target distribution
Sampling type	Random	Used `train_test_split(..., random_state=42)` without stratification
Outliers removal	No	No filtering, capping, or visual outlier check performed
Check for duplicates	No	No check using `duplicated()` or similar

Imputation of missing values	none	Checked missing values with `isnull().sum()` but applied no imputation
Drop columns	No	No explicit dropping of irrelevant columns
Encoding	none	No encoding performed; selected features were all numerical
Create new columns	No	No derived or added features
Feature selection	Yes	Manually selected columns 3 to 6 from dataset as features
Data scaling/standardisation	No	No scaler or standardisation method used
Hyperparameter tuning	No	Linear regression used with default parameters; no tuning attempted

2nd Prompt
Accuracy 11/11

Data Wrangling Step	**Technique Used**	**Details**
Check for balanced data	No	No check on distribution of target variable (`Y house price of unit area`)
Sampling type	Random	Used `train_test_split` with `random_state=42`
Outliers removal	No	No IQR, Z-score, or clipping techniques applied
Check for duplicates	No	No call to `.duplicated()` or `.drop_duplicates()`
Imputation of missing values	None	Checked with `.isnull().sum()` but no handling method used
Drop columns	No	No column was dropped from the dataset
Encoding	None	No encoding of categorical variables observed
Create new columns	No	No column creation or feature engineering done
Feature selection	No	Only selected index slices of columns (iloc[:, 3:7]) without analysis

Data scaling/standardisation	No	Features used as-is; no scaler applied
Hyperparameter tuning	No	Linear regression used without tuning; no cross-validation or search

3rd Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check via `value_counts` or visual plot of the target variable.
Sampling type	Random	Used `train_test_split(..., random_state=42)` without stratification or oversampling.
Outlier removal	No	No IQR, Z-score, or capping methods used to treat outliers.
Check for duplicates	No	No duplicate check performed using `duplicated()` or similar.
Imputation of missing values	No	`isnull().sum()` was used to inspect missing values, but no imputation was done.
Drop columns	Yes	Feature selection was manual by slicing specific columns from the dataset (using `.iloc[:, 3:7]`).
Encoding	None	No categorical features or encoding used.
Create new columns	No	No new columns derived or added.
Feature selection	Yes	Selected subset of columns manually (`iloc[:, 3:7]`) rather than using all.
Standardization	No	No scaling or normalization (e.g., `StandardScaler`) applied to features.
Hyperparameter tuning	No	Used `LinearRegression()` without any tuning or cross-validation.

4th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No countplot, histogram, or value_counts on target variable
Sampling type	Random	`train_test_split(..., random_state=42)` used without stratification
Outliers removal	No	No IQR/Z-score filtering or capping applied
Check for duplicates	No	No check using `duplicated()` or similar
Imputation of missing values	none	Missing values checked via `isnull().sum()`, but no imputation or row/column drop used
Drop columns	No	No columns dropped from dataset
Encoding	none	No categorical columns encoded
Create new columns	No	No new features created or derived
Feature selection	Yes	Only columns 3 to 6 (subset of features) used explicitly as X
Data scaling/standardisation	No	No scaler (StandardScaler, MinMax, etc.) used
Hyperparameter tuning	No	Linear regression used without tuning or search

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit check (e.g., value_counts, histplot) on target variable (y) distribution.
Sampling type	Random	`train_test_split` used without `stratify` → random sampling.
Outliers removal	No	No treatment or visualisation of outliers was performed.
Check for duplicates	No	No check for duplicates using `duplicated()` or equivalent.

Imputation of missing values	none	`.isnull().sum()` confirms no missing values; no imputation performed.
Drop columns	No	No columns were explicitly dropped.
Encoding	none	No categorical features involved; hence no encoding used.
Create new columns	No	Only slicing columns; no feature engineering or column creation.
Feature selection	No	Feature selection not performed; 4 numeric columns used directly from dataset without any correlation or importance-based filtering.
Data scaling/standardisation	No	No `StandardScaler` or other scaling/normalisation technique applied.
Hyperparameter tuning	No	`LinearRegression` used with default parameters; no tuning or search strategy applied.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit distribution check on the target (y); only missingness check and plotting of predictions.
Sampling type	Random	`train_test_split(X, y, test_size=0.3, random_state=42)` without `stratify`.
Outliers removal	No	No IQR/quantile/z-score filtering or row deletion applied.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	none	Missingness inspected via `data.isnull().sum()`, but no `fillna`/imputer or NA row/column drops.
Drop columns	No	Features selected via positional slicing (`data.iloc[:, 3:7]`) rather than dropping columns from the dataframe.

Encoding	none	No `LabelEncoder`, `OneHotEncoder`, or `pd.get_dummies()` used.
Create new columns	No	No new/engineered features created.
Feature selection	No	No correlation/variance/model-importance pruning or post-EDA drops; fixed column slice only.
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied.
Hyperparameter tuning	No	`LinearRegression` with default parameters; no `GridSearchCV`/RandomizedSearchCV/Optuna.

FOURTH RECIPE
1st Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No `value_counts()` or visualisation of target distribution
Sampling type	Random	Used `train_test_split(..., random_state=50)` without stratification
Outliers removal	No	Visualised distribution using `histplot`, but no filtering or clipping applied
Check for duplicates	No	No `duplicated()` check or removal
Imputation of missing values	none	No `isnull()` check or imputation logic
Drop columns	No	Renamed columns, but none were removed
Encoding	none	All selected features were numeric; no encoding applied
Create new columns	No	No derived features created
Feature selection	Yes	Selected 4 columns for prediction manually
Data scaling/standardisation	No	No scaling applied to features

Hyperparameter tuning	No	Linear regression used directly without tuning

2nd Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check on the target variable's distribution
Sampling type	Random	Used `train_test_split` with `random_state=50`
Outliers removal	No	No clipping, IQR, Z-score, or similar techniques used
Check for duplicates	No	No duplicate checks performed
Imputation of missing values	None	No `.isnull()` check or imputation applied
Drop columns	No	No column dropped from the dataset
Encoding	None	No encoding of categorical features
Create new columns	No	Columns renamed, but no new columns created
Feature selection	No	Manually selected 4 features without statistical criteria
Data scaling/standardisation	No	No scaler used
Hyperparameter tuning	No	Linear regression used as-is; no tuning or validation loop

3rd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value count or visualisation of target distribution.

Sampling type	Random	Used `train_test_split(..., random_state=50)` without stratification.
Outlier removal	No	Histograms plotted, but no outlier removal or treatment using IQR/Z-score.
Check for duplicates	No	No use of `duplicated()` or equivalent to check for duplicate rows.
Imputation of missing values	No	No check or handling of missing values (`isnull()` not used).
Drop columns	No	All columns retained except implicit selection for predictors.
Encoding	None	No categorical variables encoded; all columns used were numeric.
Create new columns	No	No creation of new features.
Feature selection	Yes	Only a subset of predictors used: `Distance to MRT`, `Stores`, `Latitude`, `Longitude`.
Standardization	No	No use of `StandardScaler` or other scaling method.
Hyperparameter tuning	No	Linear regression used directly without tuning.

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No countplot, histogram, or value_counts for the target variable
Sampling type	Random	`train_test_split(..., random_state=50)` used without stratification
Outliers removal	No	Visualisation (histograms) shown, but no IQR or Z-score based filtering or capping applied
Check for duplicates	No	No duplicated() check or similar operation
Imputation of missing values	none	No missing value check or imputation performed

Drop columns	No	No column was dropped from the dataset
Encoding	none	No categorical encoding applied (rename used only for readability)
Create new columns	No	No new columns created; all predictors taken directly from raw dataset
Feature selection	Yes	Only 4 predictor columns selected manually for modeling
Data scaling/standardisation	No	No scaling applied (e.g. StandardScaler, MinMaxScaler, etc.)
Hyperparameter tuning	No	Linear regression used directly with default settings

5th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value_counts or histogram on the target variable (`House price of unit area`) observed.
Sampling type	Random	`train_test_split` used without `stratify` → default random sampling.
Outliers removal	No	Histograms were plotted for distributions but no statistical or programmatic outlier handling was done.
Check for duplicates	No	No duplicate check using `duplicated()` or similar function present.
Imputation of missing values	none	No null-checking or imputation steps observed; assumed no missing values.
Drop columns	No	No columns were dropped.
Encoding	none	No categorical variables used; all columns are numeric.
Create new columns	No	No derived or engineered features were created.

Feature selection	No	A subset of features was manually selected, but no data-driven selection (correlation, importance, etc.) was applied.
Data scaling/standardisation	No	No use of `StandardScaler` or equivalent; raw features used directly in regression.
Hyperparameter tuning	No	`LinearRegression` used with default parameters; no tuning or validation strategy applied.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	Regression target; only histograms plotted (no class-balance check on a label/target).
Sampling type	Random	`train_test_split(..., test_size=0.2, random_state=50)` without `stratify`.
Outliers removal	No	No IQR/quantile/z-score filtering or row deletion applied.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	none	No `fillna`/imputer or NA row/column drops.
Drop columns	No	Columns not dropped from the dataframe; a predictor subset is selected for modeling.
Encoding	none	No `LabelEncoder`, `OneHotEncoder`, or `pd.get_dummies()` used.
Create new columns	No	No engineered features created (only renaming and selection).
Feature selection	Yes	Post-EDA manual subset for modeling: `['Distance to the nearest MRT station','Number of convenience stores','Latitude','Longitude']`.

Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied.
Hyperparameter tuning	No	`LinearRegression` with defaults; no `GridSearchCV`/`RandomizedSearchCV`/Optuna.

FIFTH RECIPE
1st Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check on target distribution using `value_counts()` or plots
Sampling type	Random	Used `train_test_split(..., random_state=42)` without stratification
Outliers removal	No	Data visualised with `pairplot` and `heatmap`, but no outlier filtering applied
Check for duplicates	No	No check using `duplicated()`
Imputation of missing values	none	No imputation used; no check like `isnull().sum()` shown
Drop columns	Yes	Dropped No column from the dataset
Encoding	none	All features were numeric; no encoding applied
Create new columns	No	No derived features created
Feature selection	Yes	Selected 6 columns manually for X; rest dropped
Data scaling/standardisation	Yes	Used `StandardScaler()` on all selected features
Hyperparameter tuning	No	Manually set learning rate and iterations for gradient descent; no formal tuning/search

2nd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check for target variable distribution
Sampling type	Random	Used `train_test_split` with `random_state=42`
Outliers removal	No	No outlier handling or clipping seen
Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` used
Imputation of missing values	None	No missing value check or imputation seen
Drop columns	Yes	Column `'No'` dropped without replacement
Encoding	None	No categorical encoding performed
Create new columns	No	No new columns derived or created
Feature selection	No	All features used; no feature ranking or correlation-based drop
Data scaling/standardisation	Yes	`StandardScaler` applied to features before training
Hyperparameter tuning	No	Gradient Descent used, but learning rate and iterations were fixed

3rd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value count or plot of target variable.
Sampling type	Random	Used `train_test_split(..., random_state=42)` without stratification or resampling.
Outlier removal	No	No use of IQR, Z-score, or capping to treat outliers.
Check for duplicates	No	No `duplicated()` or related methods used.

Imputation of missing values	No	No check or imputation for missing values (`isnull()` not used).
Drop columns	Yes	Dropped `"No"` column.
Encoding	None	No categorical encoding used; all features were numeric.
Create new columns	No	No derived or newly created columns.
Feature selection	No	Used all columns except dropped `"No"` column.
Standardization	Yes	Applied `StandardScaler` to all feature columns.
Hyperparameter tuning	No	Gradient Descent implemented from scratch; no hyperparameter tuning frameworks like GridSearch used.

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No target distribution check (e.g. countplot, value_counts)
Sampling type	Random	`train_test_split(..., random_state=42)` used without stratify
Outliers removal	No	No outlier detection or handling performed
Check for duplicates	No	No duplicated() or similar used
Imputation of missing values	none	Missing values not imputed, no dropna or fillna used
Drop columns	Yes	Only `"No"` column dropped
Encoding	none	No categorical features present or encoded
Create new columns	No	No new feature creation; only standardization and manual prediction loop
Feature selection	No	All features except the dropped ID were used

Data scaling/standardisation	Yes	`StandardScaler` applied before training
Hyperparameter tuning	No	Learning rate and iterations for gradient descent were manually set, not tuned

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value_counts or histogram check on the target variable (`Y house price of unit area`).
Sampling type	Random	`train_test_split` used without `stratify`.
Outliers removal	No	No handling or visualization of outliers, even though distribution was plotted.
Check for duplicates	No	No check using `duplicated()` or similar method.
Imputation of missing values	none	No imputation methods used; assumed no missing values based on `.info()` (no `.isnull().sum()` call).
Drop columns	Yes	Only column `'No'` was dropped
Encoding	none	No encoding applied; all features are numerical.
Create new columns	No	No new feature engineering or derived columns created.
Feature selection	No	All non-ID columns were retained; no correlation-based or model-based selection applied.
Data scaling/standardisation	Yes	`StandardScaler` applied to features before modeling.
Hyperparameter tuning	No	Gradient descent implemented manually with fixed learning rate and iterations; no parameter search or tuning strategy applied.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit distribution check of the target (`y = 'Y house price of unit area'`); EDA shows `describe`/`info`/`corr`, heatmap, and pairplot only.
Sampling type	Random	`train_test_split(X_scaled, y, test_size=0.2, random_state=42)` without `stratify`.
Outliers removal	No	No IQR/quantile/z-score filtering or row deletion; only correlation/plots.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()` in the notebook.
Imputation of missing values	none	Missingness not imputed; no `fillna`, imputer, or NA row/column drops.
Drop columns	Yes	Identifier column No dropped with `df.drop(columns="No", inplace=True)` (not reused elsewhere).
Encoding	none	No `LabelEncoder`, `OneHotEncoder`, or `pd.get_dummies()` used; features selected as numeric columns directly.
Create new columns	No	No truly new features added; scaling creates arrays (`X_scaled`) but not new columns in the dataframe (derived transformations don't count as "new").
Feature selection	No	Fixed feature subset defined upfront (X columns) without correlation/variance/model-importance pruning or post-EDA drops.
Data scaling/standardisation	Yes	`StandardScaler()` fit on `X`, producing `X_scaled` prior to splitting.
Hyperparameter tuning	No	Custom gradient descent implementation; no `GridSearchCV`/`RandomizedSearchCV`/Optuna.

SIXTH RECIPE

1st Prompt

Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No `value_counts()` or histogram on target variable shown
Sampling type	Random	Used `train_test_split(..., random_state=3)` without stratification
Outliers removal	No	Scatterplots visualised but no IQR/Z-score filtering or clipping used
Check for duplicates	No	No check using `duplicated()` or similar method
Imputation of missing values	none	No missing value check or imputation applied
Drop columns	Yes	Dropped column `'No'` explicitly
Encoding	none	No encoding step shown; all features were numerical
Create new columns	No	No derived or engineered columns
Feature selection	Yes	Dropped only the target variable, used all other columns as features
Data scaling/standardisation	No	No scaling applied to the feature set
Hyperparameter tuning	No	Linear regression used with default settings

2nd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check on target variable distribution performed
Sampling type	Random	Used `train_test_split` with `random_state=3`
Outliers removal	No	No outlier handling or clipping seen

Check for duplicates	No	No use of `.duplicated()` or related functions
Imputation of missing values	None	No `.isnull()` check or imputation shown
Drop columns	Yes	Column `'No'` dropped without replacement
Encoding	None	No categorical encoding used
Create new columns	No	No derived columns added
Feature selection	No	All features retained; no feature elimination or correlation-based drop
Data scaling/standardisation	No	No use of scalers or standardization techniques
Hyperparameter tuning	No	Linear regression used as-is without parameter search

3rd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check performed using `value_counts` or visual inspection of target distribution.
Sampling type	Random	Used `train_test_split(..., random_state=3)` without stratification or oversampling.
Outlier removal	No	Visualized outliers via scatterplots but no treatment applied.
Check for duplicates	No	No use of `duplicated()` or related methods.
Imputation of missing values	No	No check or handling of missing values.
Drop columns	Yes	Dropped `'No'` column explicitly.
Encoding	None	No encoding was required or applied (all features numeric).
Create new columns	No	No new features derived.

Feature selection	No	All columns except the target used as features.
Standardization	No	No scaling or standardization used (`StandardScaler` imported but not applied).
Hyperparameter tuning	No	Linear Regression model used with default parameters; no tuning or CV applied.

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check using value_counts or histogram on target variable
Sampling type	Random	`train_test_split(..., random_state=3)` used without stratify
Outliers removal	No	Scatterplots shown, but no filtering or capping applied
Check for duplicates	No	No call to `duplicated()` or similar
Imputation of missing values	none	No imputation or handling of missing values
Drop columns	Yes	Only `"No"` column dropped
Encoding	none	No categorical variable present or encoded
Create new columns	No	No new columns were derived or created
Feature selection	No	All columns used except target and dropped ID column
Data scaling/standardisation	No	No scaler used (e.g. StandardScaler, MinMaxScaler)
Hyperparameter tuning	No	LinearRegression used without hyperparameter tuning

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No histogram, value_counts, or similar check on target variable (`Y house price of unit area`).
Sampling type	Random	`train_test_split` used without `stratify`.
Outliers removal	No	Visualised via scatterplots, but no IQR, Z-score, or capping used for removal or adjustment.
Check for duplicates	No	No use of `duplicated()` or equivalent function.
Imputation of missing values	none	No `isnull()` check or imputation strategy applied; assumed clean dataset.
Drop columns	Yes	Column `'No'` dropped during early preprocessing
Encoding	none	No categorical variables involved, and hence, no encoding was applied.
Create new columns	No	No new features or transformations created.
Feature selection	No	All features retained except the dropped target and ID column; no correlation or model-based selection.
Data scaling/standardisation	No	No `StandardScaler` or similar technique applied; features used in raw form.
Hyperparameter tuning	No	`LinearRegression` used with default parameters; no tuning or model comparison implemented.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	Only scatterplots of features vs target; no `value_counts`/countplot/hist on the target itself.
Sampling type	Random	`train_test_split(X, y, test_size=0.2, random_state=3)` with no `stratify`.

Outliers removal	No	No IQR/quantile/z-score filtering or row deletion present.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()` in code.
Imputation of missing values	none	No `fillna`, imputer, or NA row/column drops applied.
Drop columns	Yes	Identifier column `'No'` dropped: `df.drop('No', inplace=True, axis=1)` (not reused elsewhere).
Encoding	none	All predictors used as numeric; no `LabelEncoder`/`OneHotEncoder`/`get_dummies`.
Create new columns	No	No truly new features created; X/y are defined by selecting existing columns.
Feature selection	No	No correlation/variance/model-importance pruning or post-EDA drops; modeling uses all predictors except target.
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied.
Hyperparameter tuning	No	`LinearRegression` fit with defaults; no `GridSearchCV`/`RandomizedSearchCV`/Optuna used.

SEVENTH RECIPE

1st Prompt

Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit check on target variable distribution
Sampling type	Random	Used `train_test_split(..., random_state=100)` without stratification
Outliers removal	No	Outliers visualised via histograms and boxplots, but no filtering/clipping performed

Check for duplicates	No	No check using `duplicated()` or similar methods
Imputation of missing values	none	No imputation or missing value checks shown
Drop columns	Yes	Dropped `'No'` column explicitly
Encoding	none	No encoding needed; all features were numerical
Create new columns	No	No feature engineering or new columns added
Feature selection	Yes	Dropped only target column; used all others for modeling
Data scaling/standardisation	Yes	Applied `StandardScaler()` via pipeline for linear regression
Hyperparameter tuning	No	Used default parameters for both `LinearRegression` and `RandomForestRegressor`

2nd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No target distribution check with `value_counts()` or plots
Sampling type	Random	Used `train_test_split` with `random_state=100`
Outliers removal	No	Skewness and boxplots plotted, but no removal or clipping applied
Check for duplicates	No	No check using `.duplicated()`
Imputation of missing values	None	No `.isnull()` check or imputation observed
Drop columns	Yes	Column `'No'` dropped without replacement
Encoding	None	No categorical features or encoding
Create new columns	No	No new features were derived

Feature selection	No	All features used; no statistical elimination or selection
Data scaling/standardisation	Yes	`StandardScaler()` used within pipeline
Hyperparameter tuning	No	Linear regression and random forest used with default parameters

3rd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No `value_counts()` or similar used on target variable.
Sampling type	Random	Used `train_test_split(..., random_state=100)` without stratification.
Outlier removal	No	Outliers visualised using boxplots and histograms, but not removed or capped.
Check for duplicates	No	No use of `duplicated()` or similar method.
Imputation of missing values	No	No missing value check or imputation.
Drop columns	Yes	Dropped `"No"` column.
Encoding	None	All features were numeric; no encoding required.
Create new columns	No	No new features were derived.
Feature selection	No	Used all features (excluding target) for modeling.
Standardization	Yes	Applied `StandardScaler` inside a pipeline with `LinearRegression`.
Hyperparameter tuning	No	Used default model parameters for both Linear Regression and Random Forest without tuning.

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No target distribution check (e.g. histogram, countplot, value_counts)
Sampling type	Random	`train_test_split(..., random_state=100)` used without `stratify`
Outliers removal	No	Boxplots and skewness analysed, but no filtering/capping performed
Check for duplicates	No	No use of `duplicated()` or similar check
Imputation of missing values	none	No imputation or handling of missing values
Drop columns	Yes	`"No"` ID column was dropped
Encoding	none	No encoding performed, as all features are numeric
Create new columns	No	No new features were created
Feature selection	No	All columns (except target and dropped ID) were retained
Data scaling/standardisation	Yes	`StandardScaler` applied within pipeline before LinearRegression
Hyperparameter tuning	No	Both models (LinearRegression and RandomForest) used without grid/random search

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value_counts, histplot, or target distribution check for `Y house price of unit area`.
Sampling type	Random	`train_test_split` used without `stratify` → random sampling.
Outliers removal	No	Skew and boxplots shown, but no IQR/Z-score clipping or programmatic handling of outliers.

Check for duplicates	No	No duplicate checking using `duplicated()` or similar functions.
Imputation of missing values	none	`.info()` used to inspect data; no missing value treatment needed or performed.
Drop columns	Yes	ID column `'No'` was dropped.
Encoding	none	No categorical variables in dataset; encoding was not needed or applied.
Create new columns	No	No new features were added or engineered.
Feature selection	No	All features used directly without any correlation-based, model-based, or variance-based selection.
Data scaling/standardisation	Yes	`StandardScaler` applied via a pipeline in the `LinearRegression` model.
Hyperparameter tuning	No	No GridSearchCV, RandomizedSearchCV, or tuning observed; default models used.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	Regression target; no explicit class-balance check on a label/target (only univariate hist/boxplots).
Sampling type	Random	`train_test_split(x, y, train_size=0.8, random_state=100)` without `stratify`.
Outliers removal	No	Only histograms/boxplots plotted; no IQR/quantile/z-score filtering or row deletion.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	none	No `fillna`, imputer, or NA row/column drops.
Drop columns	Yes	Identifier column `'No'` dropped via `df.drop(['No'], axis='columns', inplace=True)` (not reused elsewhere).

Encoding	none	All predictors numeric; no `LabelEncoder`/`OneHotEncoder`/`pd.get_dummies()`.
Create new columns	No	No engineered features added; only train/test splits and model pipelines.
Feature selection	No	No correlation/variance/model-importance pruning or post-EDA drops.
Data scaling/standardisation	Yes	`Pipeline([('Scaler', StandardScaler()), ('regression', LinearRegression())])` applied before fitting.
Hyperparameter tuning	No	Models trained with defaults; no `GridSearchCV`/`RandomizedSearchCV`/Optuna.

EIGHTH RECIPE
1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check like `value_counts()` or histogram on target variable
Sampling type	Random	Used `train_test_split(..., test_size=0.1)` without stratification
Outliers removal	No	No visualisation or filtering of outliers
Check for duplicates	No	No `duplicated()` check or handling shown
Imputation of missing values	none	No missing value check or imputation logic present
Drop columns	Yes	Dropped the `'No'` column explicitly
Encoding	none	All features used were numerical; no encoding needed
Create new columns	No	No derived features or new columns created
Feature selection	Yes	Used all columns except last as features (`iloc[:,:-1]`)

Data scaling/standardisation	No	No `StandardScaler` or similar applied
Hyperparameter tuning	No	Used `LinearRegression` with default settings

2nd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check on target variable distribution
Sampling type	Random	Used `train_test_split` (aliased as `tts`) without stratification
Outliers removal	No	No clipping, IQR, or z-score technique applied
Check for duplicates	No	No duplicate check performed
Imputation of missing values	None	No missing value check or imputation observed
Drop columns	Yes	`'No'` column dropped without reuse
Encoding	None	No categorical encoding observed
Create new columns	No	No derived columns created
Feature selection	No	All features used; no correlation-based or importance-based drop
Data scaling/standardisation	No	No scaling technique used
Hyperparameter tuning	No	Linear regression used with default parameters

3rd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details

Check for balanced data	No	No distribution check of target variable.
Sampling type	Random	Used `train_test_split(test_size=0.1)` without stratification.
Outlier removal	No	No outlier handling applied.
Check for duplicates	No	No use of `duplicated()` or equivalent method.
Imputation of missing values	No	No missing value check or imputation performed.
Drop columns	Yes	Column `"No"` was dropped.
Encoding	None	All features were numerical; no encoding required.
Create new columns	No	No new features were created.
Feature selection	No	All features (except target) were retained.
Standardization	No	No feature scaling or standardization used.
Hyperparameter tuning	No	Used default `LinearRegression` without tuning.

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check for target variable distribution using histogram or value_counts
Sampling type	Random	`train_test_split(test_size=0.1)` used without `stratify`
Outliers removal	No	No detection or treatment of outliers
Check for duplicates	No	No use of `duplicated()` or similar check
Imputation of missing values	none	No missing value handling or imputation performed
Drop columns	Yes	ID column `'No'` was dropped

Encoding	none	No encoding performed; all features were numeric
Create new columns	No	No new features created or engineered
Feature selection	No	All available columns except ID and target were used
Data scaling/standardisation	No	No scaling or standardisation applied
Hyperparameter tuning	No	LinearRegression used without any tuning

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No histogram, value_counts, or other distribution check on target variable (`Y house price of unit area`).
Sampling type	Random	`train_test_split` used without `stratify`.
Outliers removal	No	No statistical or visual outlier detection or treatment applied.
Check for duplicates	No	No call to `duplicated()` or similar function to identify duplicate rows.
Imputation of missing values	none	`.info()` called, but no missing value treatment performed. Assumed clean dataset.
Drop columns	Yes	ID column `'No'` dropped.
Encoding	none	No categorical variables in the dataset; encoding not required or applied.
Create new columns	No	No derived or new features were created.
Feature selection	No	All input features were used directly; no pruning based on correlation or model-based importance.
Data scaling/standardisation	No	No use of `StandardScaler` or similar preprocessing method observed.

Hyperparameter tuning	No	`LinearRegression` used with default parameters; no tuning method (grid/random search) implemented.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit distribution check on the target; only loading/EDA calls (`head/shape/columns/info`).
Sampling type	Random	`train_test_split(..., test_size=0.1)` without `stratify`.
Outliers removal	No	No IQR/quantile/z-score filtering or row deletion applied.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	none	No `fillna`, imputer, or NA row/column drops.
Drop columns	Yes	Identifier column dropped with `del df['No']` (not reused elsewhere).
Encoding	none	No `LabelEncoder`, `OneHotEncoder`, or `pd.get_dummies()` used.
Create new columns	No	No truly new features created; x/y are derived selections.
Feature selection	No	No correlation/variance/model-importance pruning or post-EDA drops.
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied.
Hyperparameter tuning	No	`LinearRegression` with defaults; no `GridSearchCV`/`RandomizedSearchCV`/Optuna.

NINTH RECIPE

1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No distribution checks (e.g., `value_counts()` or histogram) for the target variable
Sampling type	Random	Used `train_test_split(..., random_state=101)` without stratification
Outliers removal	No	Visualised residuals and scatterplots but no outlier filtering applied
Check for duplicates	No	No `duplicated()` check or removal
Imputation of missing values	none	No `isnull()` check or imputation logic shown
Drop columns	No	No columns dropped, not even `'No'`
Encoding	none	All features used were numerical; no encoding needed
Create new columns	No	No new or derived columns created
Feature selection	Yes	All columns except target used as features
Data scaling/standardisation	No	No scaling (e.g., `StandardScaler`) applied
Hyperparameter tuning	No	Linear regression used with default parameters

2nd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check on target variable distribution
Sampling type	Random	Used `train_test_split` with `random_state=101`
Outliers removal	No	No clipping or outlier filtering shown
Check for duplicates	No	No check using `.duplicated()` or similar

Imputation of missing values	None	No missing value check or imputation applied
Drop columns	No	Target column separated, but no columns were dropped fully
Encoding	None	No categorical encoding used
Create new columns	No	No derived or engineered features added
Feature selection	No	All features included in model; no elimination or importance-based drop
Data scaling/standardisation	No	No scaler or standardisation applied
Hyperparameter tuning	No	Linear regression used with default settings

3rd Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No visual or statistical check on label distribution.
Sampling type	Random	Used `train_test_split(..., random_state=101)` with no stratification or oversampling.
Outlier removal	No	Residuals plotted for analysis, but no outlier treatment was applied.
Check for duplicates	No	No code for checking duplicate entries.
Imputation of missing values	No	No imputation or missing value handling observed.
Drop columns	Yes	Dropped `"Y house price of unit area"` during splitting; `"No"` was not explicitly dropped but likely included by index.
Encoding	None	No categorical columns or encoding techniques used.
Create new columns	No	No feature engineering or new column creation.

Feature selection	No	Used all columns except the target for model training.
Standardization	No	No scaling method like `StandardScaler` applied.
Hyperparameter tuning	No	Used `LinearRegression()` directly without tuning or cross-validation.

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No countplot, histogram, or value_counts on the target variable
Sampling type	Random	`train_test_split(..., random_state=101)` used without `stratify`
Outliers removal	No	Residual analysis shown but no filtering or capping applied
Check for duplicates	No	No use of `duplicated()` or related methods
Imputation of missing values	none	No imputation or missing value handling performed
Drop columns	No	Only the target column was separated; ID column not explicitly dropped
Encoding	none	No categorical encoding needed or performed
Create new columns	No	No new columns created in the feature matrix
Feature selection	No	All columns used as predictors except the target
Data scaling/standardisation	No	No scaling (e.g. StandardScaler) applied
Hyperparameter tuning	No	Linear regression used without any hyperparameter tuning

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No target distribution check (histogram/value_counts) for `Y house price of unit area`.
Sampling type	Random	`train_test_split` used without `stratify`.
Outliers removal	No	Residuals were visualized post-prediction, but no preprocessing or outlier filtering was performed.
Check for duplicates	No	No check for duplicates using `duplicated()` or equivalent method.
Imputation of missing values	none	`.info()` was checked, but no handling of missing values was required or performed.
Drop columns	No	Only the target column was separated, and `'No'` column was not dropped; retained throughout.
Encoding	none	Dataset contained only numeric features; no encoding needed or applied.
Create new columns	No	No derived or engineered features were introduced.
Feature selection	No	All features used as-is without selection or pruning.
Data scaling/standardisation	No	No scaling (e.g. StandardScaler) applied before training.
Hyperparameter tuning	No	`LinearRegression` used with default settings; no tuning or cross-validation implemented.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit distribution check of the target (`Y house price of unit area`); only pairplot, residual plot, and histogram of residuals.
Sampling type	Random	`train_test_split(X, y, test_size=0.3, random_state=101)` without `stratify`.

Outliers removal	No	No IQR/quantile/z-score filtering or row deletion applied.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	none	No `fillna`, imputer, or NA row/column drops.
Drop columns	No	Only the target is removed to form `X`; no other columns are dropped from the dataset.
Encoding	none	No `LabelEncoder`, `OneHotEncoder`, or `pd.get_dummies()`.
Create new columns	No	No engineered features created.
Feature selection	No	All predictors (except the target) are used; no correlation/variance/model-importance pruning or post-EDA drops.
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied.
Hyperparameter tuning	No	`LinearRegression` with defaults; no `GridSearchCV`/`RandomizedSearchCV`/Optuna.

TENTH RECIPE
1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No distribution checks (e.g., histogram or `value_counts()`) on target variable
Sampling type	Random	Used `train_test_split(..., random_state=101)` without `stratify`
Outliers removal	No	No visualisation or filtering of outliers shown
Check for duplicates	No	No check using `duplicated()` or removal
Imputation of missing values	none	Confirmed in comment: “There are no null entries”; no imputation done

Drop columns	Yes	Dropped the `'No'` column explicitly
Encoding	none	No categorical columns or encoding applied
Create new columns	No	No feature engineering or new variables created
Feature selection	Yes	All columns except target used directly as features
Data scaling/standardisation	No	No use of scaling methods like `StandardScaler()`
Hyperparameter tuning	No	Linear regression used without tuning

2nd Prompt
Accuracy 11/11

Data Wrangling Step	**Technique Used**	**Details**
Check for balanced data	No	No target variable distribution check performed
Sampling type	Random	Used `train_test_split` with `random_state=101`
Outliers removal	No	No IQR, Z-score, or clipping applied
Check for duplicates	No	No `.duplicated()` or similar check shown
Imputation of missing values	None	Code comments state no nulls; no imputation methods applied
Drop columns	Yes	Column `'No'` dropped without reuse
Encoding	None	No categorical encoding used
Create new columns	No	No new columns were created
Feature selection	No	All columns except target were used; no statistical selection applied
Data scaling/standardisation	No	Model trained on raw features without scaling
Hyperparameter tuning	No	Linear regression trained without parameter tuning

3rd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No use of `value_counts` or visualisation on target variable.
Sampling type	Random	Used `train_test_split(..., random_state=101)` without stratification.
Outlier removal	No	No explicit treatment for outliers in any features.
Check for duplicates	No	No code to check for duplicates.
Imputation of missing values	No	Explicitly confirmed that there are no missing values; thus, no imputation.
Drop columns	Yes	Dropped the `"No"` column.
Encoding	None	All features were numeric; encoding not required.
Create new columns	No	No new features were created.
Feature selection	No	All columns except target were used as features.
Standardization	No	No feature scaling or standardisation applied.
Hyperparameter tuning	No	Used default `LinearRegression` model; no tuning done.

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No countplot, histogram, or value_counts used to inspect target distribution
Sampling type	Random	`train_test_split(..., random_state=101)` used without `stratify`
Outliers removal	No	No outlier detection or treatment applied
Check for duplicates	No	No call to `duplicated()` or similar

Imputation of missing values	none	Comment says “no null entries,” but no imputation or check is performed
Drop columns	Yes	ID column `'No'` dropped.
Encoding	none	No categorical columns or encodings used
Create new columns	No	No new columns created or derived
Feature selection	No	All features except ID and target used directly
Data scaling/standardisation	No	No scaler (StandardScaler, etc.) used
Hyperparameter tuning	No	Linear regression used without any tuning

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No histogram or value_counts used to inspect the distribution of the target variable.
Sampling type	Random	`train_test_split` used without `stratify`.
Outliers removal	No	No IQR, Z-score, or visual outlier detection or handling performed.
Check for duplicates	No	No `duplicated()` or similar function applied.
Imputation of missing values	none	`.info()` shows no null values, and no imputation was performed.
Drop columns	Yes	`'No'` column was dropped
Encoding	none	Dataset consists of only numerical features; no encoding required or applied.
Create new columns	No	No feature engineering or creation of new columns observed.
Feature selection	No	All input features used; no feature was pruned based on correlation or model analysis.

Data scaling/standardisation	No	No scaling (e.g., StandardScaler) applied to features.
Hyperparameter tuning	No	`LinearRegression` used with default parameters; no tuning or search strategy applied.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit distribution check of the target; only `DataFrame.info()` and notes about nulls.
Sampling type	Random	`train_test_split(X, y, train_size=0.6, random_state=101)` without `stratify`.
Outliers removal	No	No IQR/quantile/z-score filtering or row deletion present in the code.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()` in the code.
Imputation of missing values	none	No imputation applied; code comments indicate no null entries.
Drop columns	Yes	Identifier column `No` dropped via `data.drop(['No'], axis=1)`.
Encoding	none	Dataset treated as numeric; no `LabelEncoder`/`OneHotEncoder` used (note states all columns are numerical).
Create new columns	No	No engineered features created; features taken directly from existing columns.
Feature selection	No	All predictors except the target used (`X = data.iloc[:, :-1]`); no correlation/model-importance pruning.
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied anywhere.

Hyperparameter tuning	No	`LinearRegression` fitted with defaults; no `GridSearchCV`/`RandomizedSearchCV`/Optuna.