Linear Regression — Observation Sheet (Aligned to IntroTc

A) Dataset

Field	Value
Dataset Name	House Prices - Advanced Regression Tech
Owner/Source	Kaggle
Version/Date	2025-09-14
Rows	1460
Columns	81
Target Variable	SalePrice
Train/Test Split	0.8 / 0.2
Random Seed	42

B) Preprocessing

Step	Choice / Counts / Notes
Standardize column names? (Y/N)	N
Duplicates removed (count)	0
Columns with missing values	Numeric:median imputation,Categorical:r
Numeric imputation (mean/media	median
Categorical imputation (most_freq	most_frequent
Encoding strategy (None/OHE/Ord	OHE
Scaling strategy (None/Standard/M	None
Outlier handling (None/IQR/Manu	None
Feature selection/dropping (list &	dropped ID column

C) EDA

C1) Univariate (feature,	
SalePrice	right-skewed

C2) Multivariate	
Strong correlation with OverallQual(0.79), GrLivArea(0.71), GarageCa	

D) Linear Regression

Features used (list)	Optimization (Normal Eqn/GD)
Optimization (Normal Eqn/GD)	sklearn LinearRegression Normal Eqn
Metrics (Test): MAE	**MSE**
21000	1200000000
Top Coefficients	**Sign (+/-)**
OverallQual	Positive
GarageCars	Positive

oMLModule)

niques

mode imputation

some outliers

rs(0.64)

Comments	
RMSE	**R^2**
34641	0.82
Magnitude	**Interpretation**
High	Higher quality increases SalePrice
Moderate	More garage parking increases SalePrice

Train R^2	**Gap (Train–Test)**
0.89	0.07
Next Coeff	**Sign**

Comment	
Magnitude	**Interpretation**

E) Multicollinearity & Assumptions	
Top VIF #1	**Top VIF #2**
20.5	18
F) Tuning (simple → complex)	
F1) Polynomial Features	**Feature used**
	GrLivArea, OverallQual
F2) Encoding Impact	**Encoding used (None/OHE/Ordinal)**
	OHE
F3) Regularization	**Type (Ridge/Lasso)**
	Ridge
F4) Cross-Validation / Learning	**CV folds**
	3
G) Final Summary	
Chosen final model	**Why? (1–2 lines)**
Ridge Regression	due to multicollinearity stability
Aligned to headings found in IntroToMLModule: Data Loading → Univariate	

Top VIF #3	**Shapiro p**
15.8	0.06

Degrees tried	**Best degree (CV)**
1–4	1
R² before	**R² after**
0.78	0.82
Alpha grid	**Best alpha**
0.1,1,10	1
R² (CV mean)	**R² (CV std)**
0.81	0.03

Top 3 drivers	**Limitations / Ethics**
OverallQual, GrLivArea, GarageCars	Target skewness; multicollinearity am

→ Multivariate → Linear Regression → Multicollinearity (VIF) → Heteroskedasticity ·

Durbin–Watson (~2)	**Breusch–Pagan p**
1.95	0.07
Best CV R^2	**Notes**
0.75	
Did it help? (Y/N)	**Notes**
Y	
R^2 (test)	**# non-zero coefs (Lasso)**
0.83	
RMSE (CV mean)	**RMSE (CV std)**
Next steps	
Try log-transform, polynomial features, and feature engineering	
→ Regularization → CV.	

Overall (Pass/Needs Work)	**Actions**
Notes	
Bias/variance notes	