How to build and test a linear regression model

Rules are helpful, but they must be combined with good judgment (gained through experience) in order to create a meaningful model.

1) **Examine and explore the data**

a. **Look at the data**. If you can examine all of it in detail that is ideal. If dataset is very large, this may be unwieldy, so look at all of data in aggregated form (what is data type, how many entries are there, etc.) and examine a random subset in detail. Make sure the data is clean (remove NaN, for example) and meaningful (e.g., if the number of customers is negative, there is a problem).

b. **Look at statistics**: Correlation matrix, seaborn plots (to check for colinearity; compare what you see with human intuition) and probability distributions. Linear regression will be useful of the data has fairly normal distribution. Look at the error vs. y_pred plot for any weirdness (for example, due to bimodal distribution).

2) **Baselining.** Look at one feature (you might get lucky and this is THE feature needed to explain your observations). You can see the fit immediately from the seaborn plot. Run statsmodels to look at R^2, etc.
3) **Expand model.** Run it with a few features (3 or 4), so it is still humanly interpretable. Plot distribution of target variable, errors vs. predicted value to see if non-linearity or heteroskedasticity are an issue. Check fit of the model. Runs statsmodel to calculate p-values and see if they correspond to your intuition.
4) **Complete model.** Run with all features and run same diagnostics. Do you need all of the features or can you discard some of them? Are the coefficients (betas) statistically significant? Make sure you understand the diagnostics to know if your p-values are trustworthy or not.
5) **Validate model.** Set up a validation scheme, preferably cross-validation if you have the computational resources as it more robust—better to have a distribution of coefficients than point estimates. Run multiple check of the results.
6) **Challenge model.** Do you need more complexity or less? Do you need a fancier model (nonlinear)? Hopefully you have an understanding of where you need to go by knowing the variance vs. bias tradeoff.
7) **Refine model.** Add or remove complexity (using a more sophisticated model or making your model more conservative by using regularization). Compare with unrefined model (error vs. y_pred, R^2, RMSE). Use regularization to tweak model (Ridge, Lasso; Lasso may eliminate coefficients that you included at first).
8) **Test model.** How does your final, refined model perform on out-of-sample data? If it does well, hurray! If not, linear regression may not be the answer and you have to move to more complex models that are harder to interpret (adding polynomial terms, etc.).