Name: Student ID:

Today's data is on house pricing in Los Angeles! We have access to data on all of the houses that sold in four cities around west LA in the course of a single month. The `LA` data frame is available in the Files section of bCourses.

1. Our eventual goal is to predict the price of a home based on a few attributes of that home. But first, create a scatterplot examining the relationship between square footage and price.

- Write the code you used below.

- Sketch the plot below.

2. Add a column onto the `LA` dataset, `log_sqft`, which measures *log* square footage. Save the new dataset back into itself.

3. Create two side-by-side scatter plots examining the relationship between square footage *squared* and price to the left, and *log* square footage and price to the right. Sketch the plots below.

4. Explain in at least two sentences whether the logged feature or the regular feature appears more suitable to predict price when using a linear model.

5. Split the data set into 2 sets: 75% in `la_train`, 25% in `la_test`. Copy the code below. *Hint: check the tutorial*

6. Fit several linear models, predicting `price` with ...

- square footage;
- square footage raised to the second power (square footage squared);
- log square footage.
- 2nd degree polynomial of square footage
- ....
- 10th degree polynomial of square footage

Write a `dplyr` pipeline for `la_train` to find the *training* RMSE, and R^2 for the model. Then do the same for `la_test` *You'll need to calculate the predictions* `price_hat`*, the residuals* `e_hat`*, then summarize the MSE, the variance of* `price`*, and finally with those building blocks, calculate RMSE and* R^2

Model	R^2_{train}	R^2_{test}	$\widehat{\text{RMSE}}_{\text{train}}$	$\widehat{\text{RMSE}}_{\text{test}}$
`price ~ sqft`				
`price ~ sqft^2`				
`price ~ poly(sqft, degree = 2, ...)`				
`price ~ poly(sqft, degree = 3, ...)`				
`price ~ poly(sqft, degree = 4, ...)`				
`price ~ poly(sqft, degree = 5, ...)`				
`price ~ poly(sqft, degree = 6, ...)`				
`price ~ poly(sqft, degree = 7, ...)`				
`price ~ poly(sqft, degree = 8, ...)`				
`price ~ poly(sqft, degree = 9, ...)`				
`price ~ poly(sqft, degree = 10, ...)`				

7. What are the units of measurement of R^2 and RMSE?

8. Which of the several models would you use? Should you use RMSE or R^2 to make this decision? And should you use test or train?

9. Are you seeing any sign of overfitting? Explain