Q1

1. Fit a simple linear regression model of `price ~ sqft.` Interpret the coefficients for the intercept and `sqft`.

To *interpret* model coefficients in this case study means: write down in concrete terms what each coefficient says about the value of a home, or the change in the value of a home, conditional on predictors. For example, suppose the coefficient for x is 1.5. To interpret this number would entail saying something like this: “An increase in 1 in x is associated with a change of 1.5 in y, on average, holding the other predictors constant.”

2. Center `sqft` and refit the model. The intercept changes a lot after centering. Why?

Note: there are several options for centering.

- Create a new centered variable in your data and use that in the model. You would simply subtract the mean of the column from every observation.
- Use the `scale()` function in the model formula: `scale(sqft, scale = F)`.(`scale = F` means: center but don’t scale. The default for the function is to center and scale.)
- Do the centering yourself in the model formula using the `I()` function: `I(sqft - mean(sqft))`.

Q2

1. Fit a multiple regression model of `price` using all the available predictors. Center the numeric predictors. Interpret the coefficients in this model (including the intercept).

Remember that Long Beach is the (missing) reference city (assuming that factor levels have been assigned alphabetically).

Q3

1. To the above model add an interaction between centered `sqft` and `city`. This means that you combine these terms multiplicatively (*) rather than additively (+).
2. Create a visualization of the interaction between `sqft` and `city`.
3. Interpret the two interaction coefficients in this model.

Interaction models can be tricky to understand. Here is some guidance:

- The **intercept** is the average value of the target when the inputs are 0 (for numeric variables) or the reference category (for categorical variables).
- The **main effects** are the non-interaction coefficients in the output. In the plot you created for this question you can see that there is a regression line for each city. Similarly in the interaction model: there is no *single* relationship between `price` and `sqft`, and consequently the main effect for `sqft` is conditional on `city`. Specifically, it denotes the relationship between `sqft` and `price` for the reference city. The same is true for the main effects of `city`: they are conditional on `sqft` = 0. The main effects for a predictor in an interaction model will always be conditional on the levels of the variable with which it has been interacted.
- The **interaction effects** are the final 2 coefficients in the output. (The colon indicates the interaction, as in `sqft:citySanta Monica.`) These coefficients estimate the *change in the slope of the regression line* for each city *compared to the reference city*. If the interaction coefficients are positive that means that the regression line relating sqft to price is steeper for that particular city in comparison to the reference city, or, equivalently, that the relationship is stronger.

Q4

Is this a good model? To assess the interaction model fit create a residual plot with model residuals on the vertical axis and the fitted values on the horizontal axis. One easy way to get a residual plot in R is to use `plot(model, which = 1)`. Comment on model fit.

Q5

What should Andrew say in his presentation? Write a brief summary of the quantitative evidence that he should use to support this recommendation.

Challenge Question 1

Are these differences between cities practically significant (as opposed to just statistically significant)? Use the information from the above regression to calculate the price returns associated with increasing the size of (pick your numbers) a 3000 square foot SFR worth 500k by 20% or 40% or…? Additionally, use the SE in the regression output to incorporate uncertainty into your answer, in order to report a range rather than a single summary dollar amount.

Challenge Question 2

Write a practice question for the Multiple Linear Regression section of *Learn R + Statistics*. It should explore the content in one of the lessons and be modeled on this question, with both a hint and an answer.