For this programming assignment, you will extend the "Example with semi-realistic data" section at the end of the "Linear regression: deep dive" notebook. Review that notebook, and go through the "Example with semi-realistic data" section, before you begin your answer to this question.

You will create a new notebook to answer this question, rather than modifying the existing notebook. Make sure to include your name and net ID in a text cell at the top of the notebook. Also, include text cells throughout the notebook in which you briefly explain each step, and note any important observations. Be as specific as you can in noting your observations.

In your new notebook,

a. Read in the "Advertising" data, and make sure it is loaded correctly. Visually inspect the data using a pairplot, and note any meaningful observations. In particular, comment on which features appear to be correlated with product sales, and which features appear to be correlated with one another. Then, split the data into training data (70%) and test data (30%). (You may copy code from the "Deep Dive" notebook, but you should write text in your own words.)

b. Use the training data to fit a simple (univariate) linear regression to predict product sales, for each of three features: TV ad budget, radio ad budget, and newspaper ad budget. In other words, you will fit three regression models, with each model being trained on one feature. For each of the three regression models, create a plot of the data and the regression line, with product sales (y) on the vertical axis and the feature on which the model was trained (x) on the horizontal axis. (You can include both the training and test data in the plots, but use one color for training data and another color for test data.) Also, for each regression model, print the intercept and coefficients, and compute the MSE and R2 on the training data, and MSE and R2 on the test data. Comment on the results. Which type of ads seems to have the greatest effect on product sales? Which regression model is most effective at predicting product sales? (You may copy code from the "Linear regression: deep dive" notebook, but you should write text in your own words.)

c. We know that computing MSE or R2 is not sufficient to diagnose a problem with a linear regression. (Refer to the "It's not just noise" section in the "Linear regression: deep dive" notebook for further discussion of this topic.) Create some additional plots as described below to help you identify any problems with the regression. You can include both test and training data in these plots, but use one color for training data and a different color for test data:

- For each of the three regression models, plot predicted sales ($\hat{y}$) on the vertical axis, and actual sales (y) on the horizontal axis. Make sure both axes use the same scale. Comment on your observations. What would you expect this plot to look like for a model that explains the data well?

- For each of the three regression models, compute the residuals $(y - \hat{y})$. Note that some of these will be negative, and some will be positive. What is the mean residual for each of the regression models? What *should* be the mean residual for a well-fitted regression model? Explain your answer.
- For each of the three regression models, plot the residuals $(y - \hat{y})$ on the vertical axis, and actual sales (y) on the horizontal axis. Use the same scale for all three subplots. Comment on your observations. Is there a pattern in the residuals (and if so, what does it indicate), or do they appear to have no pattern with respect to actual sales?
- For each of the three regression models AND each of the three features, plot the residuals $(y - \hat{y})$ on the vertical axis, and the feature (x) on the horizontal axis. This plot will include nine subplots in total. Make sure to clearly label each axis, and also label each subplot with a title that indicates which regression model it uses. Is there a pattern in the residuals (and if so, what does it indicate), or do they appear to have no pattern with respect to each of the three features?

d. Next, fit a multiple linear regression to predict product sales, using all three features to train a single model: TV ad budget, radio ad budget, and newspaper ad budget. Print the intercept and coefficients, and compute the MSE and R2 on the training data, and MSE and R2 on the test data. Comment on the results. Make sure to explain any differences between the coefficients of the multiple regression model, and the coefficients of the three simple linear regression models. (You may copy code from the "Linear regression: deep dive" notebook, but you should write text in your own words.) Also repeat the analysis of part (c) for this regression model:

- Plot predicted sales $(\hat{y})$ on the vertical axis, and actual sales (y) on the horizontal axis. Make sure both axes use the same scale. Comment on your observations. What would you expect this plot to look like for a model that explains the data well?
- Compute the residuals $(y - \hat{y})$. What is the mean of the residuals? What *should* be the mean of the residuals for a well-fitted regression model? Explain your answer.
- Plot the residuals $(y - \hat{y})$ on the vertical axis, and actual sales (y) on the horizontal axis. Comment on your observations. Is there a pattern in the residuals (and if so, what does it indicate), or do they appear to have no pattern with respect to actual sales?
- For each of the three features, plot the residuals $(y - \hat{y})$ on the vertical axis, and the feature (x) on the horizontal axis. Make sure to clearly label each axis. Is there a pattern in the residuals (and if so, what does it indicate), or do they appear to have no pattern with respect to each of the three features?

e. Our multiple linear regression includes additive effects of all three types of advertising media. However, it does not include *interaction* effects, in which combining different types of advertising media together results in a

bigger boost in sales than just the additive effect of the individual media. The pattern in the residuals plots from parts (a) through (d) suggest that a model including an interaction effect may explain sales data better than a model including additive effects. Add four columns to your data frame: `newspaper` $\times$ `radio`, `TV` $\times$ `radio`, `newspaper` $\times$ `TV`, and `newspaper` $\times$ `radio` $\times$ `TV`. Then, train a linear regression model on all seven features: the three types of ad budgets, and the four interaction effects. Repeat the analysis of part (d) for the model including interaction effects. Comment on the results. Are the interaction effects helpful for explaining the effect of ads on product sales? Are there any patterns evident in the residual plots that suggest further opportunities for improving the model?