PDAT 610: Module 6 Homework

Introduction

This assignment revisits the Ames Housing data set one more time. So, to get started, make sure to load the data set as we've done before!

Submission Method

- Create an R Markdown document containing both your code and any written responses or explanations.
- Make sure to number your answers so that the grader can easily see which question you were answering.
- The code for any question part should be contained in a single code block
- When you're done, knit your code to a PDF file, and submit that file through Blackboard.
- Hint: Sometimes students only test their code by running the individual chunks interactively in R Studio. I'd suggest checking often to make sure your code will knit to a final document. It's easier to diagnose any knitting problems if you know exactly what you just typed that caused the knitting to break (as opposed to waiting until the end and troubleshooting the whole document).

Assignment

1. To start this assignment, let's take a look at the relationship between a home's overall condition (`Overall_Cond`) and its sale price.

    a. Use ggplot to make a boxplot that illustrates the relationship between `Sale_Price` and `Overall_Cond`.

    b. Describe the general relationship between overall condition and sale price. How does this relationship break down for the "Average" category?

2. Let's explore what might be leading to the breaking of the pattern for the "Average" condition group.

    a. It might be the case that the way home condition is recorded is not the same for older homes and newer homes. That could especially be the case if the ratings were not all done at the same time. On the other hand, perhaps some average-condition homes sell for more if they're on bigger lots.

    Using ggplot, make a jitter plot (`geom_jitter`) with overall condition on the *x* axis, sale price on the *y* axis, color mapped to the year the home was built, and size mapped to the lot area. [Note: A *jitter plot* is a scatter plot where an extra random term is added to the *x* and *y* coordinates to prevent dots from exactly overlapping.]

    Your graph should show good attention to detail. That might mean using scale or theme to (at least)

    - include human-readable axis labels and titles (not the default variable names),
    - include appropriate legends, and
    - make sure that labels don't overlap, etc.

b. Briefly describe what your graph shows you about the relationship of `Year_Built` and `Lot_Area` to `Sale_Price`, especially among the "average" condition homes. In other words, which seems to have the stronger relationship to `Sale_Price`?

c. In part (b), you should have identified either `Year_Built` or `Lot_Area` as being related to the high median price of "average" condition homes.

   In a sense, this explains why the "average" category doesn't look like the others, but it doesn't explain why the variable you identified in part (b) seems to be related to higher sale prices. Perhaps there is a lurking variable (or more than one).

   For this part, choose one other variable that you believe will be highly related to `Sale_Price`, and create a graph that displays `Sale_Price`, the variable you identified in (b), and the new explanatory variable you've identified in part (c). Then use `ggplot` to create a graph with appropriate mappings to display the relationship between all three variables. Make sure to pay attention to graphical detail, as you did in part (a).

   [Hint: In choosing an appropriate graph, you might go back to the beginning of the lecture slides, choose a good graph form for displaying two variables, then add a third mapping to that. Or you might experiment until you find something that really works for your data!]

d. Now let's sum it up. Write a short paragraph that explains your hypothesis from part (c), describes what your graph showed about the relationship between the three variables, and concludes with whether your graph seems to be consistent with your hypothesis from part (c).