

Week 7. Homework 6. R-Data Visualization and Regression

These questions are designed to help you practice and reinforce your understanding of different visualization and regression techniques using R programming.

Submission Instructions

At the end of the assignment, ensure to submit your work as a Markdown (.Rmd) file. Name your file as follows: HW6_YourName.Rmd, replacing "YourName" with your actual name. Make sure to include both the code and its output in the Markdown file for full credit. Upload your .Rmd file to the designated submission folder on the course portal by the due date.

Deadline: Make sure to submit your homework by next Friday at 1pm.

House_Price Dataset Specification

- ID: Unique identifier for each house.
- Location: Categorical variable indicating the location zone (e.g., Urban, Suburban, Rural).
- Size_sqm: Size of the house in square meters.
- Bedrooms: Number of bedrooms.
- Bathrooms: Number of bathrooms.
- Age_years: Age of the house in years.
- Garage: Indicates if the house has a garage (Yes, No).
- Condition: Overall condition (e.g., Excellent, Good, Fair, Needs Work).
- Price: Price of the house in USD.
- Distance_to_city_center_km: Distance to the nearest city center in kilometers.

Questions

1. Use ggplot2 to create a scatter plot of Size_sqm versus Price. Color the points by Location.
2. Plot Age_years versus Price and add a regression line to the plot. How does the age of the house appear to affect its price?
3. Modify the scatterplot of Size_sqm versus Price by setting the graphical parameters to use larger point sizes and a different color for points. Use theme() to adjust the plot theme.

4. Create a bar chart showing the average house price for each Location. Use `geom_bar(stat="identity")` with pre-summarized data.
5. Plot a histogram of Price. Adjust the bins to appropriately display the distribution.
6. Overlay a density plot on top of the histogram for Size_sqm. Adjust the transparency of the histogram for better visualization.
7. Create a bar plot that shows the count of houses by Condition. Customize the colors of the bars.
8. Use `facet_wrap()` to create a scatter plot of Size_sqm versus Price for each Location in separate panels.
9. Perform a linear regression with Price as the dependent variable and Size_sqm, Bedrooms, Bathrooms, and Age_years as independent variables. Use `lm()` and display the summary of the model.
10. Extend the regression analysis by including Distance_to_city_center_km as an additional predictor. Check the summary and interpret the coefficients. How does distance to the city center impact house prices?