1 Collect and prepare data

Download the data from Kaggle. Prepare the data for analysis and make your own selection of variables to analyze, keeping in mind the below exercises. There is a large number of continuous and discrete variables in the data set, which lend themselves naturally to regression analyses. Convert the categorical variables in your analyses to type *factor*, before usage. Present summary statistics of the data set with the variables of your choice using function *stargazer* and present a concise description of the main characteristics of your variables. Use function *ggplot* to make three plots of your choice where in each plot the dependent variable *SalesPrice* is related to another variable. For subsequent reporting, it is advised to use function *ggsave* to export these figures in *pdf* or *png* format. The choice of appropriate plot types, e.g., scatter plot, bar chart, etc., is part of the exercise. [0.5-1 pages]

2 Theoretical model and OLS assumptions

Develop a theory to explain variation in the main dependent quantity *SalesPrice*. Present a causal relationship diagram, and formulate three research hypotheses as illustrated in class. Formulate the population regression model, including appropriate specification of categorical variables, interactions and non-linearities (if present). Elaborate on each of the six linear regression model assumptions discussed in class and explain why which assumptions could be violated; use appropriate example variables to explain why you think which assumptions might be violated. [0.5-1 pages]

3 OLS regression and model fit

Estimate the model developed in the preceding exercise, both with and without interaction and non-linear terms. Report your results using tables generated with function *stargazer*. Present and interpret the estimation results. Also determine which of the independent variables in your model have the larger effect sizes; motivate the criterion you used. [1-1.5 pages]

4 Diagnostic checking

Perform a diagnostic analysis of the estimated model, i.e., systematically analyze if OLS model assumptions have been violated, and if remedies for these violations seriously affect model outcomes. For instance, check if inference is robust for transformations of the explanatory variables and for heteroskedastic standard errors. Also, analyze the (standardized) residuals to assess whether their behavior is consistent with the assumptions of the linear regression model, and check if multicollinearity problems adversely affect the interpretation of the estimation results. Re-run updated models to determine if any remedies or adjustments have been effective to counter the observed modeling issues; discuss consequences of your model interventions by comparing the results with the originally estimated models. [0.75-1.25 pages]

5 Subset analyses

The last exercise is about the consequences of sub-sample analyses for your findings. Sub-samples can be defined based on existing categorical variables or discretized quantitative variables of your liking. Analyze whether differences exist between the estimation results obtained for your model in the different sub-samples. Run different models to deliver insights into meaningful subsets of your data. Present your results in tables, and concisely discuss your findings. Pay special attention to accurately describing the different groups you analyze and how the results change between your groups. Please present additional figures clarifying the differences you might find. [1-1.5 pages]