Applied Linear Statistical Models
Assignment 5

This assignment is in one part. **Please work in groups of two or three**. Write the name and email of your group members on your solution. It is sufficient to hand in one solution per group. You should work in a `.Rmd` file and turn in the resulting `.html` or `.pdf` file. **Please turn in your solution before 23:59 on October 25th (Monday)**.

Part I (100%): Modeling

This assignment is long so start early and show up to workshops. **Remember that this assignment will always count towards your final grade and it has a weight of 20%**.

The data

You will use data that was used to estimate the value of real estates in Iceland for 2017. The data `gagnasafn_endurmat2017_litid.csv` can be found in Canvas. There are 34865 real estates and 22 variables in the data set. A description of the variables can be seen at the end of this document. A map of `matssvaedi` can be found in Canvas. Further description of the data can be found in `Fasteignamat2017.pdf` in Canvas.

The assignment

You should build a model that can be used to predict the value of real estates (`nuvirdi`) in the following neighborhoods in Reykjavik (one model for all five neighborhoods): (i) `Miðbær frá Bræðraborgarstíg að Tjörn`; (ii) `Melar að sjó`; (iii) `Háleiti/Skeifa`; (iv) `Hólar, Berg`; (v) `Réttarholt`. Filter the data for the appropriate neighborhoods.

Select a seed based on the day of the date of birth of one of your group members. For example if a member of your group was born July 1st 1964, the seed should be 1. Split your data into a training data set and a test data set. The size of the test data should not exceed 30% of your full data.

Perform an exploratory analysis on your training data. Compute means, standard deviations, quantiles. Find minimums and maximums. Check variable correlations and if there are any missing values. Make sure that categorical variables are indeed categorical and not numerical. Describe your findings.

The goal is to fit a model which minimizes RMSE,

$$RMSE := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

of your training data without overfitting. To do this, you should compute the RMSE of your model for the training data **and** your test data. This process requires multiple iterations of fitting, diagnosing, and transforming. Create the following table and fill in the values for each of your iterations.

Model	#Predictors	R^2_{adj}	AIC	RMSE *training*	RMSE *test*
i	p	x	y	z	t

You should include the diagnostics of the **first model you fit** and the **last model you fit**. To bridge the two models you should describe the in-between models. Refer to the table above as necessary. Once you are satisfied with your RMSE, write out the equation of your final model and show estimates of all parameters in the model along with an estimate for σ.

You are free to do whatever you want **so long as you justify your actions**. Throw out outliers (or not), create now variables (or not), use splines (or not), use polynomials (or not), Box-Cox transformations of the response (or not) and so on. Note, if you transform y the RMSE will be on that transformed scale. Be sure to compute the RMSE on the original scale. Remember that there is no, one single model correct here. All models are wrong but some are useful.

Variable	**Description**
rfastnum	ID of real estate
kdagur	Date of purchase
nuvirdi	Value of real estate
teg_eign	Type of property
svfn	ID of municipality
byggar	Year of construction
haednr	Floor number
lyfta	Elevator available (1 yes, 0 no)
ibm2	Square meters of property
fjhaed	Number of floors included
fjbilast	Number of parking spaces included
fjbkar	Number of bath tubs included
fjsturt	Number of showers included
fjklos	Number of toilets included
fjeld	Number of kitchens included
fjherb	Number of rooms included
fjstof	Number of living rooms included
fjgeym	Number of storage spaces included
stig10	Measurement of building completeness (10 fully built)
matssvaedi	Location of property
undirmatssvaedi	Sub-location of property
ibteg	Type of property (numerical)