Multiple Lineaer Regression on Ames Housing Data

Stat 632

Zhaoshan Duan

April 21, 2021

Contents

1 Abstract

The real estate market is often seen as an important reflection of the economy. Knowing common factors that influence the housing prices is of great interests for sellers and buyers. In this project, we are primarily interested in predicting sale prices of residential properties given explanatory variables that describe different aspects of the properties with a Multiple Linear Regression model.

We will apply Multiple Linear Regression technique

Univariate multiple regression

plotting

exploring regression models to answer scientific questions

presenting the results

find a good model

2 Problem and Motivation

2.1 Background of the Dataset

This project uses the Ames Housing Dataset by De Cock (2011). A contemporary alternative to the well known Boston Housing dataset. This dataset describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The original dataset contains 2930 observations and 82 explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involving assessing home values. Most of the explanatory variables are information that a home buyer typically sought out when purchasing properties.

The continuous numeric variables relate to the various area dimensions for each property and the discrete numeric variables quantify the number of typical items within the house such as number of bedrooms. The nominal categorical variables identify types of dwellings, garages, materials and environmental conditions while ordinal categorical variables rate items within the property. Dean De Cock (2011)

2.2 Motivation

For many people, homeownership is a both a dream and an achievement. It is a serious purchasing decision that requires meticulous research and careful Pro & Con analysis. Accurate prediction of housing prices provides great help in buyer's decision-making process and informs them what characteristics of the properties generally affect the prices. The same can be said for realtors. sellers and developers. Housing prices also reflect the health of the economy, which can be insightful for police makers.

While there are many external factors influencing the housing price of a given property such as crime rate in the region, proximity to public schools, hospitals and busy areas, fixed characteristics of the house is often the first thing people look at. Therefore, we think it will be interesting to predict the sale price of the property using a multiple linear regression model with explanatory variables that contain information about many common aspects of the house such as number of bathrooms, size of the basement and so on. It would also be fascinating to observe the quantitative relationship between these characteristics and the sale price, and see which ones have the most influence as well as its implication to buyers' purchasing behavior.

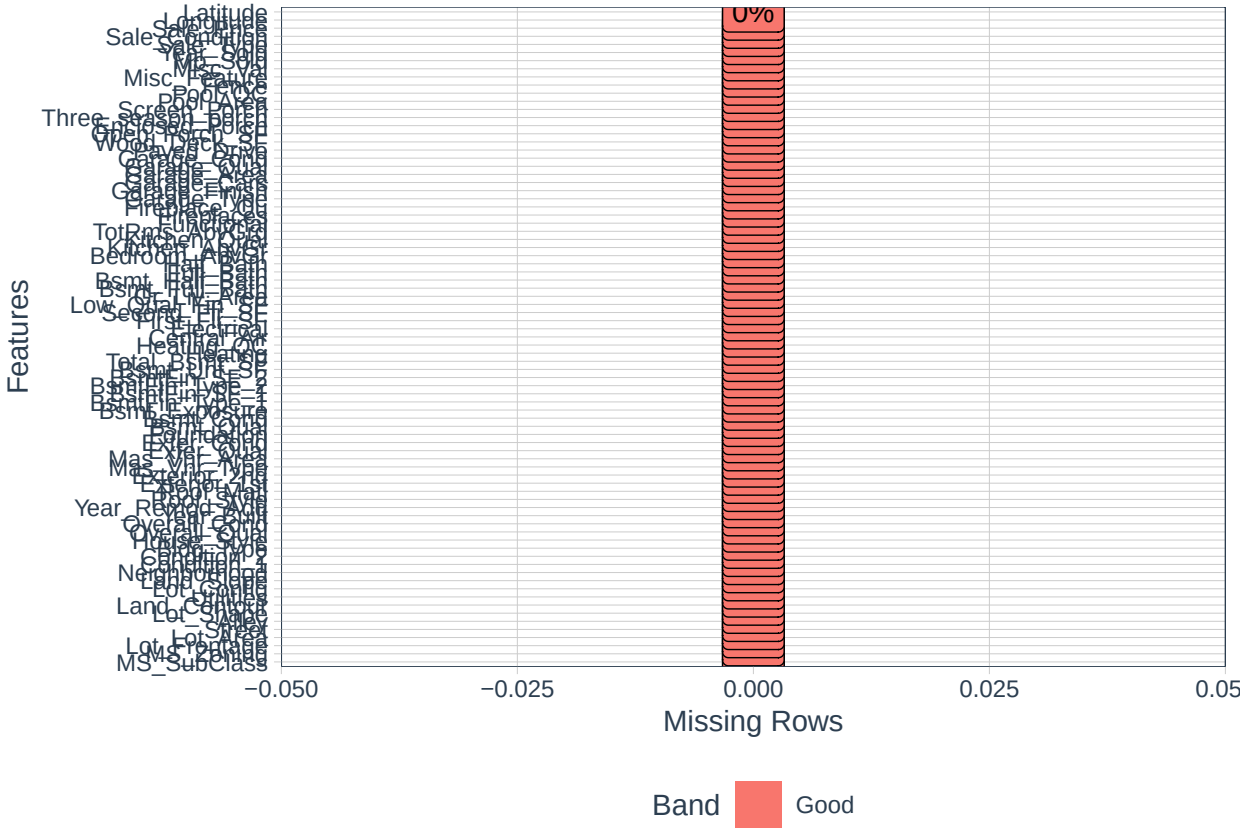
3 Data Description

We use the `AmesHousing` package on CRAN to access the data. The package provides two version of the data: `ames_raw` and processed `ames`. Our processing and analysis are done on the processed version as it removes unique identifiers such as `Order` and `PID` , arranges all factors unordered, and engineers features with large missing values. This results in a dataset with 2930 observations and 81 explanatory variables.

3.1 Exploratory Data Analysis

3.1.1 Missingness

We first investigate the missing values of the dataset using `DataExplorer` package. All the missing values have already been removed and recoded since this project is using the processed version of the data.

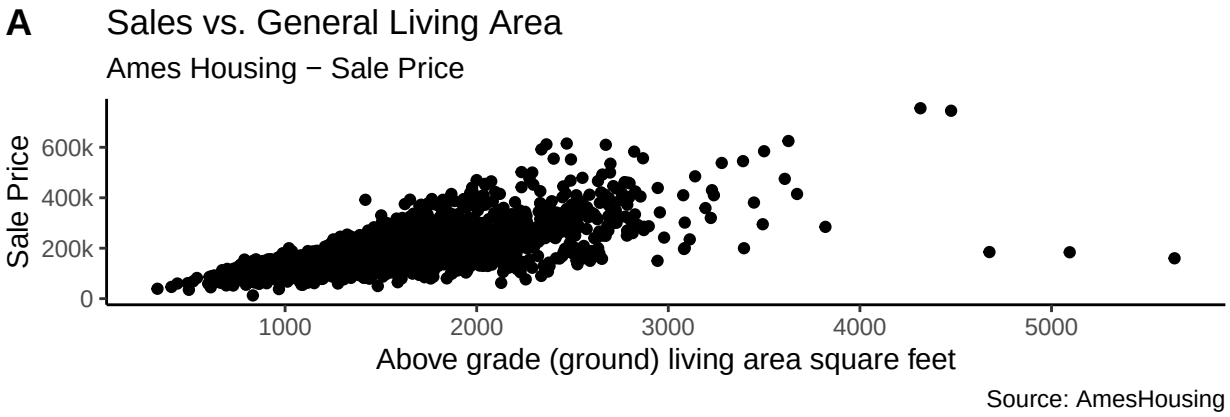


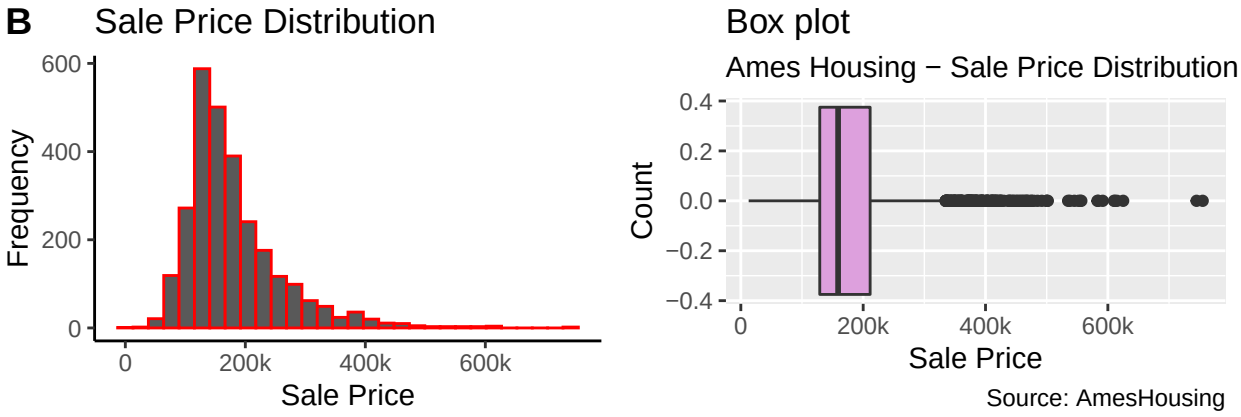
3.1.2 Summary Statistics

We then look at the summary statistics of the response variable `Sale_Price`, numeric features and categorical features respectively.

3.1.2.1 Response Variable `Sale_Price` By examining the descriptive statistics of the target variable, `Sale_Price`, we learn that:

- `Sale_Price` has a mean value of $179,957.7
- `Sale_Price` has a median value of $159,000

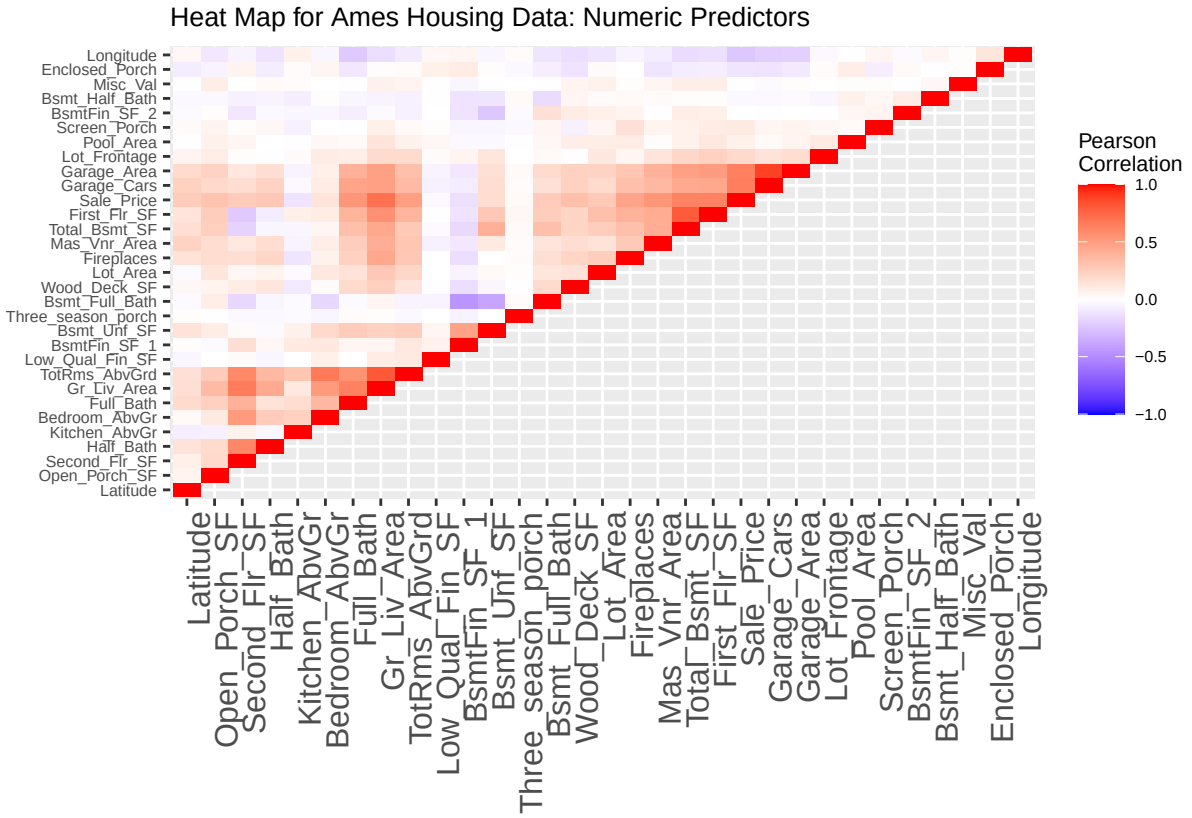




`Sale_Price` Statistics	Values
Mean	1.799577×10^5
Median	1.59×10^5

3.1.3 Correlation

We start by visualizing the correlation coefficients matrix of the numeric variables using heat map. Some correlation between the predictors are obvious and intuitive such as, generally `Gv_Liv_Area`: Above grade (ground) living area in square feet, and should be directly associated with, `First_Flr_SF`: First Floor living area in square feet, and `Bedroom_AbvGr` : Area of Bedrooms above grade.



3.1.4 Anomalous Analysis

We use the `anomalize` package to

3.1.5 Summary Statistics

Because our research focus is on traditional residential houses, we removed 25 observations that are classified as commercial properties, and 139 observation that are floating village properties, 2 observations that are industrial properties, and 2 agricultural properties. This leave us with a dataset with 2762 observations.

Table 2: MS_Zoning Summary

MS_Zoning	n
Floating_Village_Residential	139
Residential_High_Density	27
Residential_Low_Density	2273
Residential_Medium_Density	462
Agriculture	2
Commercial	25

MS_Zoning	n
Industrial	2

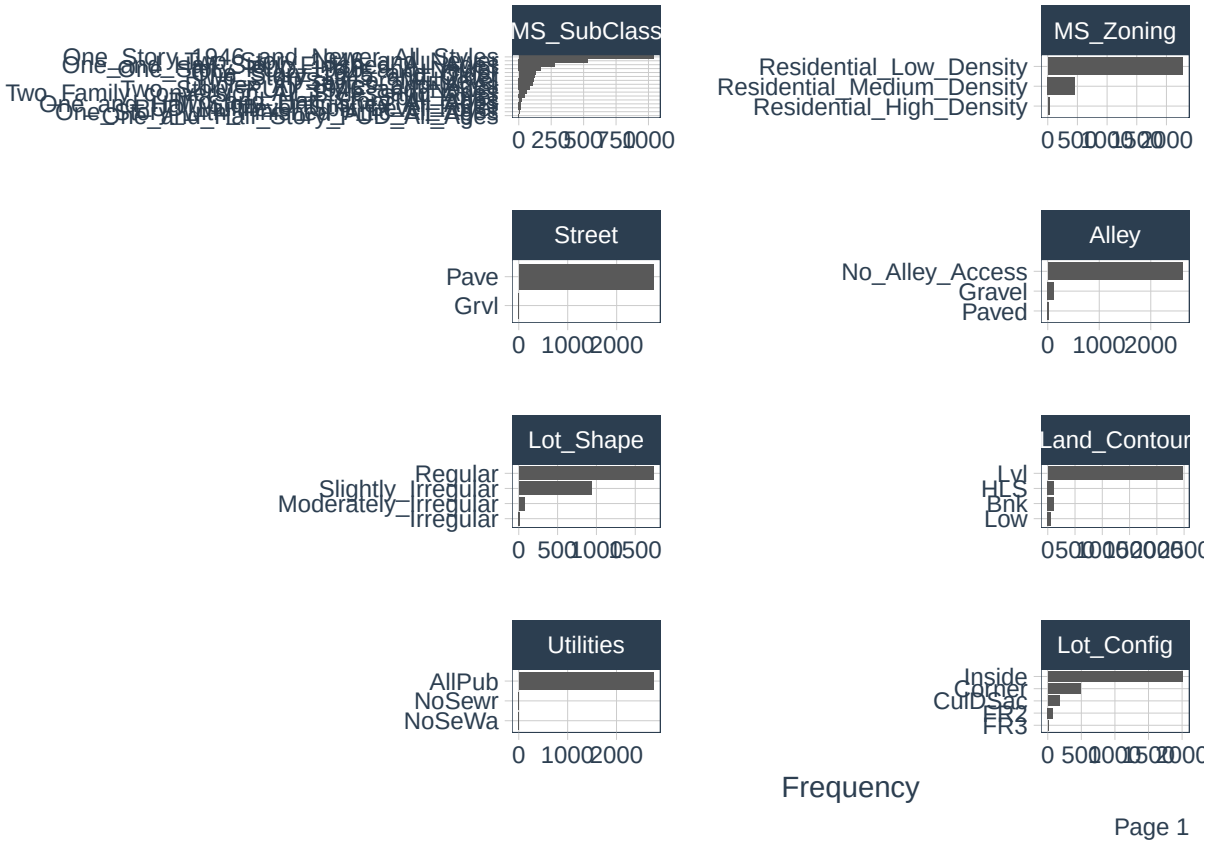
Ames dataset fits into the long-established *hedonic pricing method* to analyzing housing prices.

neighborhood-specific and unit-specific characteristics help determine house prices.

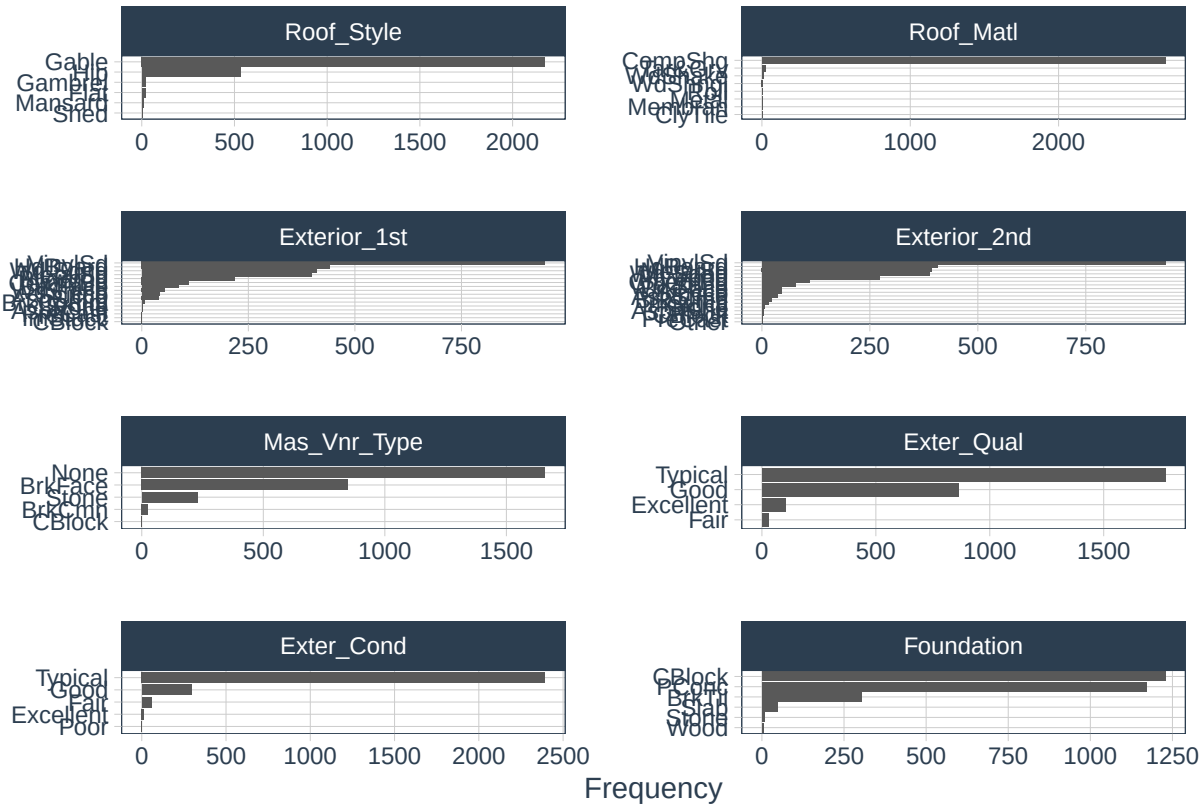
describe the relevant variables that you will use in your project.

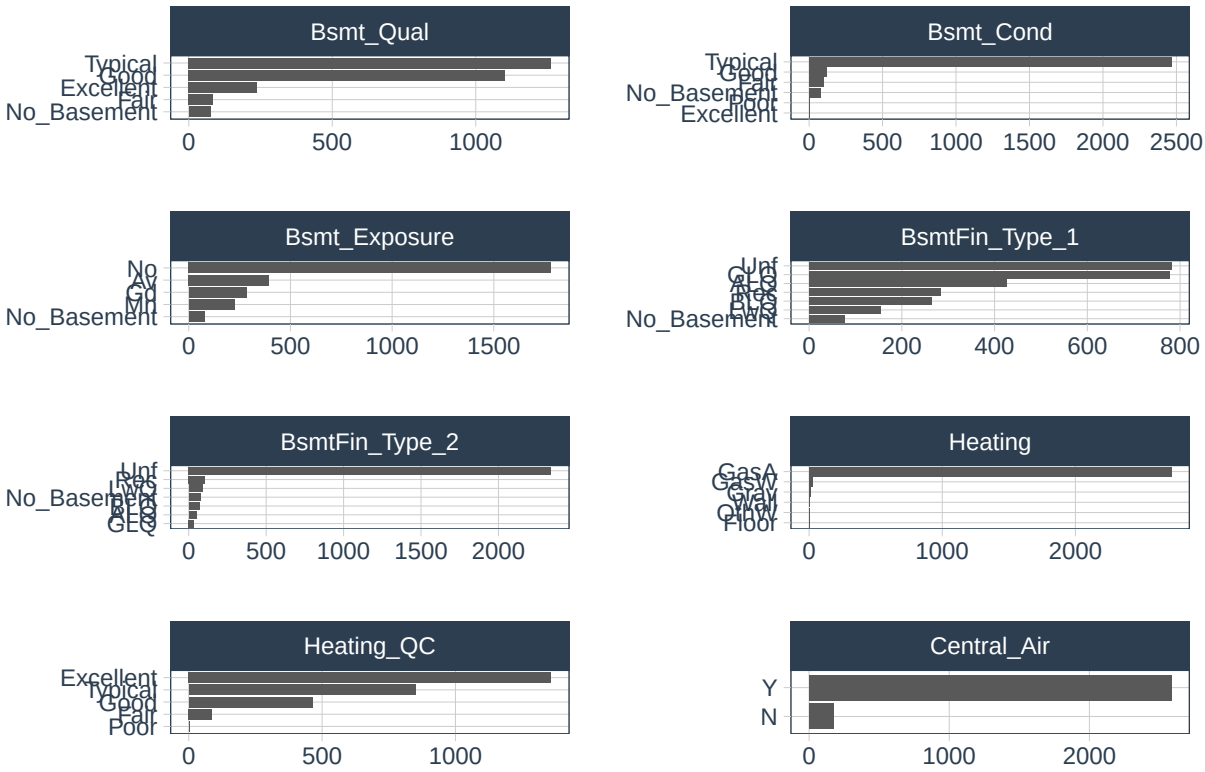
In this section, also present and discuss relevant summary statistics and graphical displays of your dataset. Be selective about the descriptive statistics that you decide to include

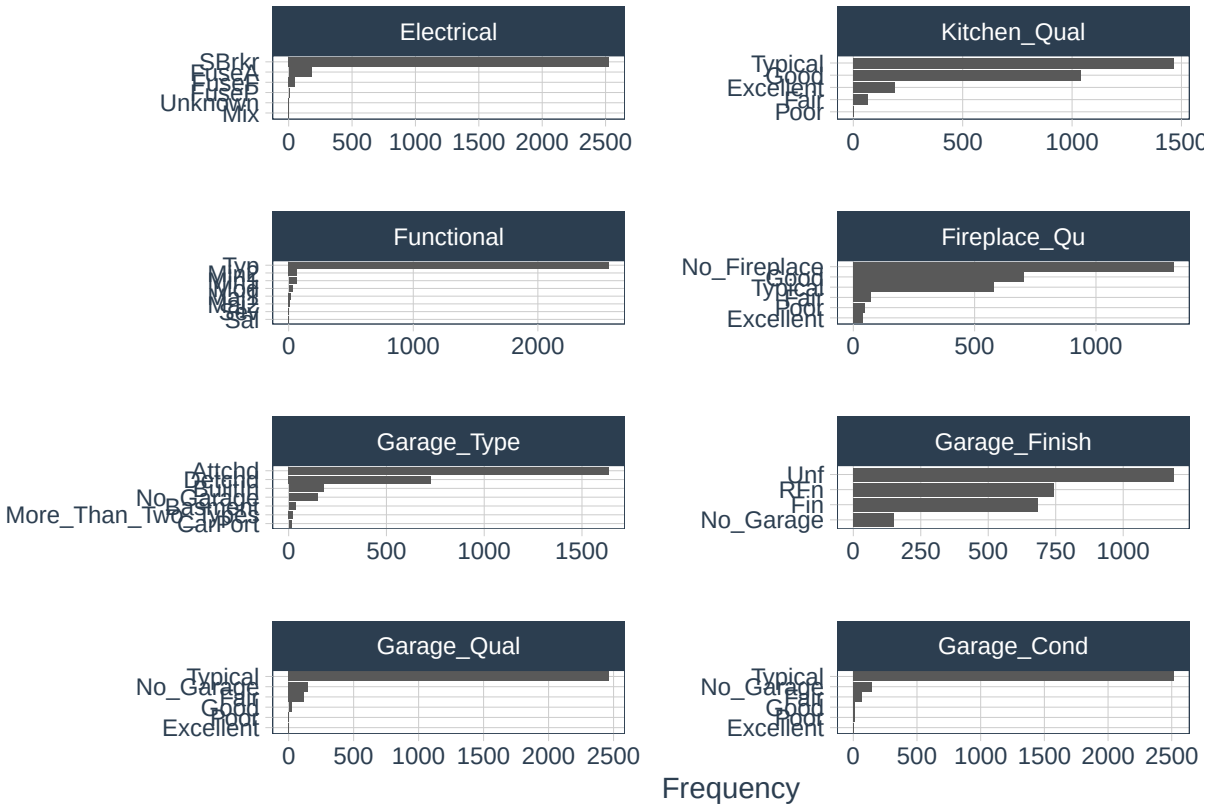
```
## 2 columns ignored with more than 50 categories.
## Year_Built: 116 categories
## Year_Remod_Add: 61 categories
```

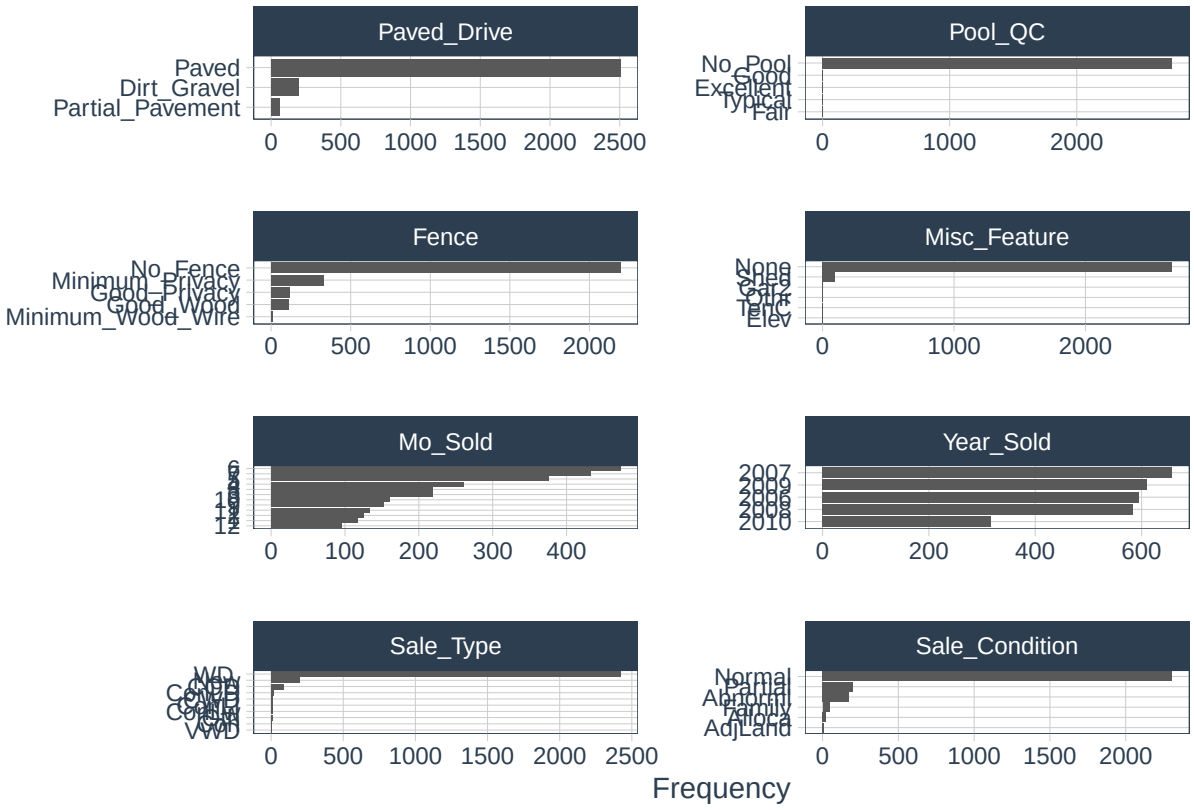



Land_Slope
Gtl
Mod
Sev
0
1000
2000
Neighborhood
South_and_West_of_Iowa_
0
100
200
300
400
Condition_1
Norm
0
500
1000
1500
2000
Condition_2
Norm
0
1000
2000
Bldg_Type
OneFam
TwnhsE
Duplex
Twnhs
TwoFmCon
0
500
1000
1500
2000
House_Style
One_Story
One_and_Half_Fin
0
500
1000
Overall_Qual
Average
Above_Average
Very_Poor
0
200
400
600
800
Overall_Cond
Average
Above_Average
Below_Average
Excellent
Very_Poor
0
500
1000
1500
Frequency


Roof_Style
Gable
Hip
Gambrel
Flat
Mansard
Shed
0
500
1000
1500
2000
Roof_Matl
CompShg
Tar&Grv
WdShake
WdShngl
Metal
Roll
Membran
ClyTile
0
1000
2000
Exterior_1st
VinylSd
0
250
500
750
Exterior_2nd
VinylSd
0
250
500
750
Mas_Vnr_Type
None
BrkFace
Stone
BrkCmn
CBlock
0
500
1000
1500
Exter_Qual
Typical
Good
Excellent
Fair
0
500
1000
1500
Exter_Cond
Typical
Good
Fair
Excellent
Poor
0
500
1000
1500
2000
2500
Foundation
CBlock
PConc
BrkTil
Slab
Stone
Wood
0
250
500
750
1000
1250
Frequency


Bsmt_Qual
Typical
Good
Excellent
Fair
No_Basement
0
500
1000
Bsmt_Cond
Typical
Good
Fair
No_Basement
Poor
Excellent
0
500
1000
1500
2000
2500
Bsmt_Exposure
No
Av
Gd
Mn
No_Basement
0
500
1000
1500
BsmtFin_Type_1
Unf
GLQ
ALQ
BLQ
Rec
LwQ
No_Basement
0
200
400
600
800
BsmtFin_Type_2
Unf
Rec
LwQ
No_Basement
BLQ
ALQ
GLQ
0
500
1000
1500
2000
Heating
GasA
GasW
Grav
Wall
OthW
Floor
0
1000
2000
Heating_QC
Excellent
Typical
Good
Fair
Poor
0
500
1000
Central_Air
Y
N
0
1000
2000
Frequency


Electrical
SBrkr
FuseA
FuseF
FuseP
Unknown
Mix
0 500 1000 1500 2000 2500
Kitchen_Qual
Typical
Good
Excellent
Fair
Poor
0 500 1000 1500
Functional
Typ
Min2
Min1
Mod
Maj1
Maj2
Sev
Sal
0 1000 2000
Fireplace_Qu
No_Fireplace
Good
Typical
Fair
Poor
Excellent
0 500 1000
Garage_Type
Attchd
Detchd
BuiltIn
Basment
No_Garage
CarPort
More_Than_Two_Types
0 500 1000 1500
Garage_Finish
Unf
RFn
Fin
No_Garage
0 250 500 750 1000
Garage_Qual
Typical
No_Garage
Fair
Good
Poor
Excellent
0 500 1000 1500 2000 2500
Garage_Cond
Typical
No_Garage
Fair
Good
Poor
Excellent
0 500 1000 1500 2000 2500
Frequency



3.1.6 Feature Selection

Variables removed from the model:

`Neighborhood` is relevant only when the interest is to model the location effect.

`MSZoning` labels commercial observations as `C`.

General living area with more than 4000 square feet has also been removed from the dataset according to the recommendation from paper

data with house sold year at 2010 has been removed since it does not cover information for the whole year

4 Question of Interest

My primary interest of research is **predicting the sale price of a residential house sale based on common, fixed characteristics of that a property agents/assessors would use to assess property value**.

predicting housing price, `Sale_Price` ,

Describe in plain English the questions that your analysis will answer. Scientific, not statistical, terminology should be used here. For example, words like 'association,' 'effect,' or 'relationship' are okay, while 'p-value,' 'coefficient,' or 'regression' are not.

5 Regression Analysis, Results and Interpretation

5.1 Model Selection

Model 1

A strong analysis should include the interpretation of the various coefficients, statistics, and plots associated with their model and the verification of any necessary assumptions

In the first model you are allowed only limited manipulations of the original data set. You are allowed to take power transformations of the original variables [square roots, logs, inverses, squares, etc.] but you are NOT allowed to create interaction variables. This means that a variable may only be used once in an equation [if you use x2 don't use x]. Additionally, you may eliminate any data points you deem unfit. This model should have a minimum r-square of 73% and contain at least 6 variables. The intent of this project is for the majority of your effort to be devoted to creating and reviewing this model.

Model 2

experiment with any of the methods that were discussed during the semester for finding better models and are allowed to create any new variables they desire (such as quadratic, interaction, or indicator variables).

evaluated through a cross-validation or data splitting technique where the original data set is split into two data sets: the training set and the validation set. The students are given the training set for the purpose of developing their model and I retain the validation set for use in evaluating their model. A relative grade is assigned by comparing their fit on the validation set to that of their fellow students with bonus points awarded to those who substantially exceed their fellow students and point reductions occurring for models which fit exceedingly poorly (See section 4 - Evaluating the Models for more details).

Training and testing

80/20

Your narrative should include • Important Details of the Analysis: – Perform the analysis in R. Depending on the questions you want to answer, this will include various items from the following list: computing coefficient estimates, R2 adj, p-values or test statistics, confidence intervals, prediction intervals, model selection procedures and results, diagnostics, etc. Do not simply use every single method we've discussed in class; you will need to convince the reader that you have used the appropriate tools for answering the question of interest. – If you did a hypothesis test, then state your null and alternative hypothesis, the value of your test-statistic, p-value, decision, and conclusion. Provide similar detail for confidence intervals, submodel tests, ANOVA tables, etc. You can pull this information from R output, but it should be stated in the text so the reader doesn't have to go looking for it. • (Exploratory Analysis) Exploratory plots of the data and numerical summaries are essential in beginning any analysis. At this stage, scatterplots, added variable plots, boxplots, etc. can give you a sense of relationships that exist between relevant variables. Will transformations be needed/useful for any of these variables? You should comment on your findings, particularly

if there are interesting or counterintuitive observations to be made. Additionally, if there is any preliminary evidence that important regression assumptions may be violated, you should mention them and suggest remedies. • Diagnostic Checks: – Were your assumptions plausible? Why? How did you check them? – The diagnostics shown in your report are used to show that the analyses are valid. In other words, there should be no blatant violation of the linear regression assumptions. Of course, it may take a while to arrive at a model with good diagnostics, requiring transformations for example. In this case, you need to demonstrate the necessity of these remedial steps, either by referring to your exploratory analysis or, most likely, by including in the Appendix (see below) diagnostic plots for preliminary models which showed violations of the assumptions. • Interpretation: – What do your results mean for the questions you were trying to answer? – All exploratory and diagnostic plots should be shown. Relevant plots with proper title, variable names, legends, etc must be included within the body of the text.

Plots should be readable, but should not take up an entire page. That means that plots should not be too small or too large. – After conducting the analysis, you should give concrete (i.e. data-specific), accurate and complete interpretations of your results. These interpretations should involve a mix of statistical terminology, variable names and appropriate scientific units. If you are using hypothesis tests, do not focus too much on p-value 0.05 or any other significance level, but rather on how strongly (or weakly) the data serve as evidence against the null hypothesis.

Heatmap

Overall F-test first

then partial F-test

outliers and high leverage points

6 Conclusion

Summary of your findings and any comments you may have about the reliability or generalizability of your analysis. In this section you should summarize your findings based on your final model in clearly understandable, non-statistical terms. What is the main message produced by your analysis? There may also be additional questions that arise, problems you encounter, or possible extensions of your analysis that could be addressed here. Also, you may include any final comments and thoughts about your project. For example, do you trust your results? How general are your results, to what situations do they apply? Any other comments.

7 Appendices

7.1 Appendix 1: R

Github link

7.2 Appendix 2:

Any exploratory data analysis from the project, or figures and plots that you found interesting, but not of primary importance to your final analysis. For example, this appendix is appropriate to show diagnostic plots, Box-Cox plots, etc., for preliminary regression models which were not used as they showed violations of model assumptions. This may not be looked at for grading but could be useful for your own future reference. All tables and figures should be numbered and referred to by number (e.g., Figure 1, Figure 2, Table 1, etc.).

8 Workflow

1. frame the problem and look at the big picture
2. get the data
3. Explore the data to gain insgihts
4. Prepare the data to better expose the underyling data ptterns to ML algorithms
5. Explore many differnt models and shortlist the best ones
6. fine-tune models nad combine them into a great solution
7. present solution
8. launch, monitor, and maintain the system

Reference

Dean De Cock. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." http://jse.amstat.org/v19n3/decock.pdf.