Assignment 1: Getting know your Data

Sri Seshadri

6/22/2017

Introduction

This report discusses the Ames, Iowa housing data set, its quality and observations from Exploratory Data Analysis (EDA). EDA was used to identify potential predictor variables that can be used for predicting value of a "typical" home.

Sample Definition

It is assumed that typical home buyers are those that move from apartments to single family or town homes. Also apartments are less likely to be sold to individuals as they remain holdings of owners for rental income. Single family and town homes belong to "Residential Low density" (RL) zoning classification in the city of Ames. Data belonging to only to the RL zone is considered for this analysis. There are a total of 2930 data points, out of which the rows belonging to the following zoning classifications are not sampled. The sampled data however has few two-family conversion and duplex houses, which are not removed from the sample at this time.

Table 1: Drop waterfall

Zoning	counts
A (agr)	2
C (all)	25
FV	139
I (all)	2
RH	27
RM	462

Data cleaning and Sanity check

From the Ames, Iowa housing dataset's dictionary (see Apendix) and based on some preliminary EDA the variables listed in Table 2 and Table 3 were chosen to be important for furhter analysis. The dataset is fairy clean and there are 2 points that are missing. Table 2 shows the summary statisics of the numeric variables and it is noted that statistics are within reasonable bounds and appear to be in the units of measure as described in the data dictionary. Table 3 shows the chosen categorical (Nominal) variables and their unique number of values, sample sixe (n) and count of missing data. 4 of the 26 neighborhood are not in the sample frame chosen.

Table 2: Data sanity check for numeric variables

	min	Q1	median	Q3	max	mean	sd	n	missing
LotArea	1700	8339	10000	12160	215245	11143.83	8428.74	2273	0
OverallCond	1	5	5	6	9	5.52	1.03	2273	0
YearRemodel	1950	1967	1993	2004	2010	1985.30	19.88	2273	0
TotalBsmtSF	0	855	1056	1367	6110	1111.84	448.40	2273	0
GrLivArea	334	1149	1479	1788	5642	1535.49	516.33	2273	0

	min	Q1	median	Q3	max	mean	sd	n	missing
BsmtFullBath	0	0	0	1	3	0.47	0.53	2272	1
BsmtHalfBath	0	0	0	0	2	0.07	0.26	2272	1
FullBath	0	1	2	2	4	1.60	0.55	2273	0
HalfBath	0	0	0	1	2	0.40	0.51	2273	0
BedroomAbvGr	0	3	3	3	6	2.91	0.79	2273	0
TotRmsAbvGrd	2	6	6	7	15	6.56	1.53	2273	0
GarageCars	0	1	2	2	4	1.85	0.74	2273	0
MoSold	1	4	6	8	12	6.22	2.72	2273	0
YrSold	2006	2007	2008	2009	2010	2007.78	1.31	2273	0
SalePrice	35000	137500	172000	223500	755000	191283.25	81295.74	2273	0

Table 3: Data sanity check for nominal variables

	# Unique	n	missing
LotConfig	5	2273	0
Neighborhood	22	2273	0
HouseStyle	8	2273	0
KitchenQual	5	2273	0
SaleCondition	6	2273	0
## Initial Explo	ratory Data	Analys	is