Exploratory Data Analysis

Introduction to exploratory data analysis

- Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

Introduction to exploratory data analysis

- Whenever we engage in a predictive modeling activity, we need to first understand the data for which we are working from. This is called Exploratory Data Analysis or EDA.

- The primary purpose for the EDA is to better understand the data we are using, how to transform the data, if necessary, and how to assess limitations and underlying assumptions inherent in the data structure.

- Data scientists need to know how the various pieces of data fit together and nuances in the underlying structures in order to decide what the best approach to the modeling task.

- Any type of method to look at data that does not include formal statistical modeling and inference generally falls under the EDA.

EDA:-

Here are some of the main reasons why we utilize EDA:

- Detection of mistakes.
- Checking of assumptions.
- Preliminary selection of appropriate models and tools.
- Determining relationships of the explanatory variables (independent).
- Detecting the direction and size of relationships between variables.

Features of a Data Set

Variables

Observations

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
$61,000.00	6360	2	1	1	yes	no	no
$66,000.00	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	3	1	2	yes	no	no

Quantitative Variables

- Numeric

Qualitative Variables

- Categorical

Dependent Variable

Independent Variables

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
$61,000.00	6360	2	1	1	yes	no	no
$66,000.00	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	3	1	2	yes	no	no

Data Munging

- Data Munging is the transformation of raw data to a useable format.
- Many datasets are not readily available for analysis.
- Data needs to be transformed or cleaned first.
- This process is often the most difficult and the most time consuming.

Data Munging Tasks

Data Munging tasks include:

- Renaming Variables
- Data Type Conversion
- Encoding, Decoding, recoding data.
- Merging Datasets
- Transforming Data
- Handling Missing Data (Imputation)
- Handling Anomalous values

- These data munging tasks are an iterate process and can occur at any stage throughout the overall EDA procedure.

Understanding the data

Non-Numeric

Missing Values

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
	6360	2	1	1	yes	no	no
	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	15	1	2	yes	7	no

Outlier

Error

Observation: The first row should contain variable names and all of the data should be completely filled after the data munging process is complete.

Data munging Tasks

Renaming Variables

- The names of variables should make intuitive sense to non-practitioners and does not have to conform to IT protocols and standards.

T1K5X
$42,000.00
$38,500.00
$49,500.00
$60,500.00

→

Price
$42,000.00
$38,500.00
$49,500.00
$60,500.00

Data Type Conversion

- Depending upon the modeling task at hand and the software, the data may need to be expressed in a specific format in order to process correctly.

Date
January 1st, 2014

→

Date
1/1/2014

SQL: Text String
Varchar (max)

SQL: Date Value
Datetime

Data munging Tasks

Data Munging Tasks

Encoding Data

- There are times when we need to change the underlying contents in a variable to prepare them for analytics. Ex. Qualitative to Quantitative.

driveway
yes
no
yes

driveway
1
0
1

- If we are using categorical variables, we need to clean them to get rid of non response categories like "I don't know", "no answer", "n/a", etc... We also need to order the encoding of categories (potentially reverse valence) to ensure that models are built and interpreted correctly.

Response
Strongly Agree
Strongly Disagree
Agree
Disagree
No Preference

Response
Strongly Agree
Agree
No Preference
Disagree
Strongly Disagree

Response
Strongly Agree
Agree
Disagree
Strongly Disagree

Response
4
3
2
1

- Usually non response categories is coded with values like 999. If this was a value in the variable "Age", this will skew the results and should be turned to NULL and reviewed further.

Data munging Tasks

Merging Datasets

- It is quite rare that you will have a dataset readily constructed for analysis. This may require some data manipulation and merging in order to get the data in the correct form.

ID	price	lotsize
A1234	$42,000.00	5850
A1235	$38,500.00	4000
A1236	$49,500.00	3060
A1237	$60,500.00	6650
A1238	$61,000.00	6360
A1239	$66,000.00	4160
A1240	$66,000.00	3880
A1241	$69,000.00	4160
A1242	$83,800.00	4800
A1243	$88,500.00	5500
A1244	$90,000.00	7200
A1245	$30,500.00	3000
A1246	$27,000.00	1700

ID	bedrooms	bathrms	stories	garagepl
A1234	3	1	2	1
A1235	2	1	1	0
A1236	3	1	1	0
A1237	3	1	2	0
A1238	2	1	1	0
A1239	3	1	1	0
A1240	3	2	2	2
A1241	3	1	3	0
A1242	3	1	1	0
A1243	3	2	4	1
A1244	3	2	1	3
A1245	2	1	1	0
A1246	3	1	2	0

ID	price	lotsize	bedrooms	bathrms	stories	garagepl
A1234	$42,000.00	5850	3	1	2	1
A1235	$38,500.00	4000	2	1	1	0
A1236	$49,500.00	3060	3	1	1	0
A1237	$60,500.00	6650	3	1	2	0
A1238	$61,000.00	6360	2	1	1	0
A1239	$66,000.00	4160	3	1	1	0
A1240	$66,000.00	3880	3	2	2	2
A1241	$69,000.00	4160	3	1	3	0
A1242	$83,800.00	4800	3	1	1	0
A1243	$88,500.00	5500	3	2	4	1
A1244	$90,000.00	7200	3	2	1	3
A1245	$30,500.00	3000	2	1	1	0
A1246	$27,000.00	1700	3	1	2	0

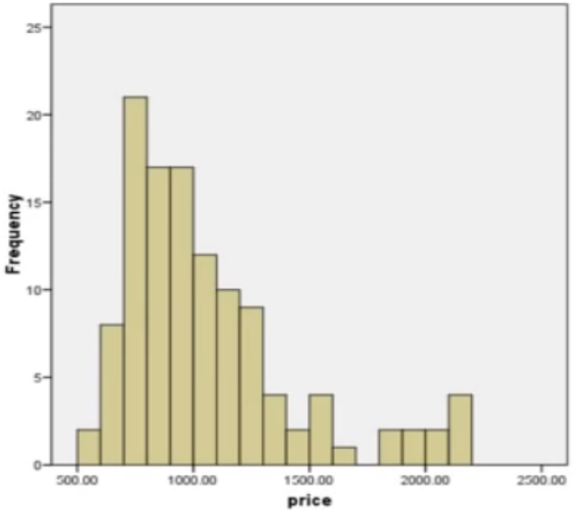
Observation: The datasets will need to have a common ID as the link to join the data. After the data has been merged, the ID may not be necessary to retain for model building.

Data munging Tasks

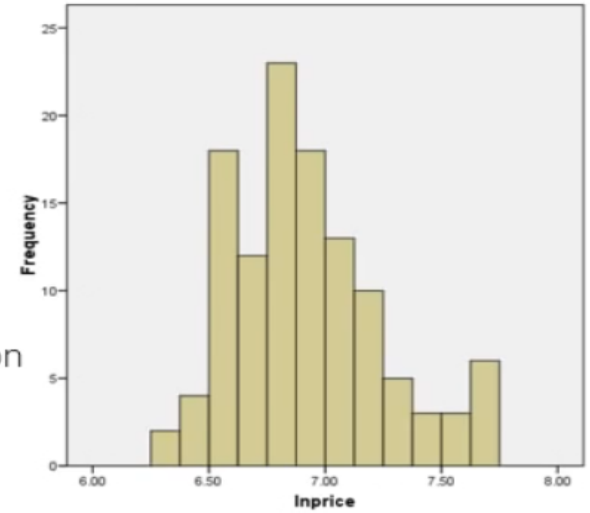
Data Munging Tasks

Transforming Variables

- There may be times where a variable will need to be transformed in order to achieve linearity. This will aid in strengthening the results of parametric based methods.



Natural Log Transformation



Data munging Tasks

Imputation

- If there are missing values in a column, these cannot be left unattended. We must decided if we want to:
  - Remove the observation from the dataset
  - Calculate a value for the null (impute). This usually is determined with the mean or median, however, a more advanced version can use a multiple linear regression formula.

price
$ 42,000.00
$ 38,500.00
$ 49,500.00
$ 60,500.00
$ 66,000.00
$ 69,000.00
$ 83,800.00
$ 88,500.00
$ 90,000.00
$ 30,500.00
$ 27,000.00

price
$42,000.00
$38,500.00
$49,500.00
$60,500.00
$58,660.00
$58,660.00
$66,000.00
$69,000.00
$83,800.00
$88,500.00
$90,000.00
$30,500.00
$27,000.00

Mean = 58,660

Median = 60,500

Data munging Tasks

Handling Anomalous Values

- Depending upon the analytic task, we need to assess points which exhibit a great deal of influence on the model.
- Outliers are data points that deviate significantly from the spread or distribution of other similar data points. These can typically be detected through the use of scatterplots.
- Many times we will delete the entry with an outlier to achieve normality in the dataset.
- In some instances, an outlier can be imputed but this must be approached with caution.

- **Important:** The drivers of outlying data points need to first be understood prior to devising an approach to dealing with them. They can hold the clues to new insights.