Business Analytics

Exploratory Data Analysis & Data Cleaning

Proschool
An IMS Initiative

Suppose we want to predict the house prices depending on some of the variables, we might need to build a regression model for prediction of house prices!!!

But before jumping to model building its preferred to study and understand the data we have.

Lets have a look at the data…

Name	Age	Gender	Education	Salary	AppraisedValue	Location	Landacres	HouseSizesqrft	Rooms	Baths	Garage
Tony	25	M	Grad	50	700	Glen Cove	0.2297	2448	8	3.5	2
Harret	52	F	PostGrad	95	364	Glen Cove	0.2192	1942	7	2.5	1
Jane	26	F	PostGrad	65	600	Glen Cove	0.163	2073	7	3	2
Rose	45	F	Grad	100	548.4	Long Beach	0.4608	2707	8	2.5	1
John	42	M	Grad	77	405.9	Long Beach	0.2549	2042		1.5	1
Mark	62	M	PostGrad	118	374.1	Glen Cove	0.229	2089	7	2	0
Bruce	51	M	Grad	101	600	Glen Cove	0.1714	1344	8	1	0
Steve	43	M	Grad	108	299	Roslyn	0.175	1120	5	1.5	0
Carol	24	F	PostGrad	51	471	Roslyn	0.213	1817	6	2	0
Henry	25	M	PostGrad	68	510.7	Roslyn	0.1377	2496		2	1
Donald	41	M	Grad	86	517.7	Long Beach	0.2497	1615	7	2	1
Maria	51	F	Grad	122	1200	Long Beach	0.4116	4067	9	4	1
Janet	49	F	PostGrad	112	700	Roslyn	0.3372	3130	8	3	1
Sophia	32	F	Grad	85	374.8	Roslyn	0.1503	1423		2	0
Jeffery	37	M	Grad	90	543	Long Beach	0.2348	1799	6	2.5	1

This dataset contains information about individuals and details about their dwellings.

Looking at the dataset we might have some questions in mind:

- What could be the average age of the people in the data?
- What is the average salary of the people?
- Why are there missing values in the rooms column and how they be replaced?
- What the general observation…are people graduates or post graduates?
- Does the salary depend on age ?
- Does the house appraised value depend on the number of rooms or area of the house or both? Can we predict house values using the given data?
- Etc…

And many more questions will follow along with the answers as we dig the data deeper and this process of mining the data is called exploratory analysis!!

What is Data Exploration

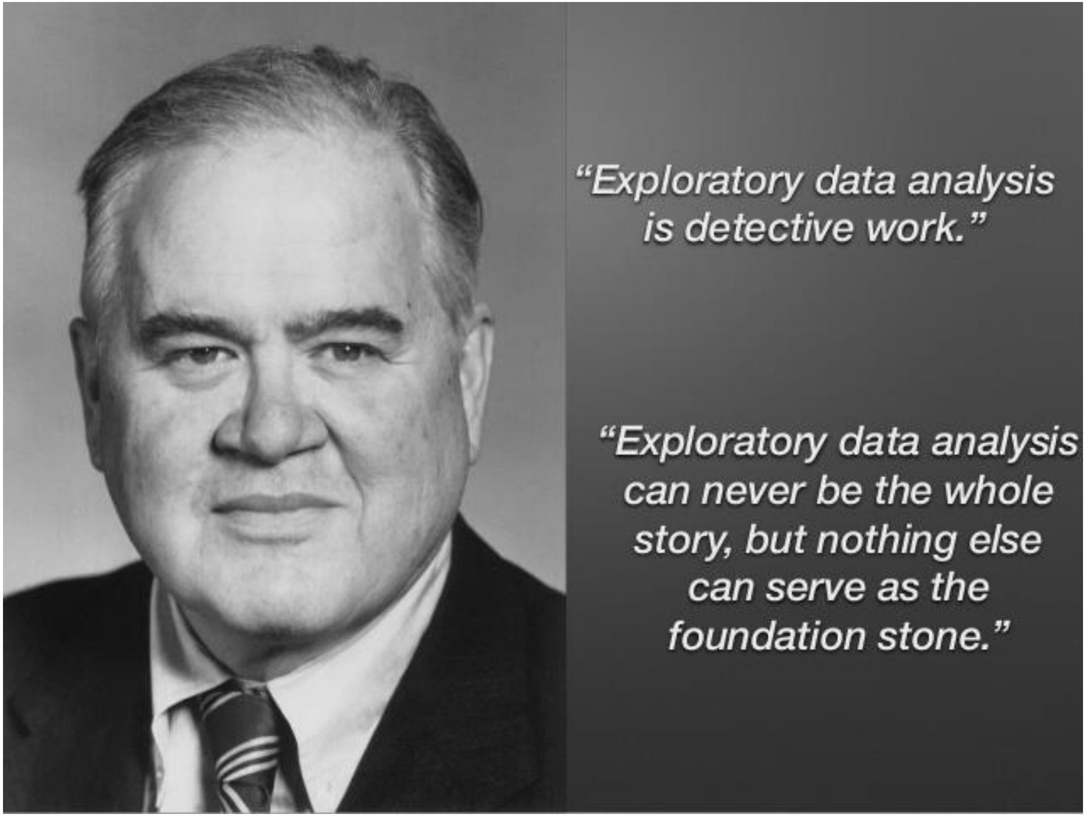
- If we wish to build an impeccable predictive model, neither any programming language nor any machine learning algorithm can award it to you unless you perform data exploration.

- Data Exploration not only uncovers the hidden trends and insights, but also allows you to take the first steps towards building a highly accurate model

- Major time needs to be spent on data exploration, cleaning and preparation as this would take major part of the project time

- Data cleaning can support better analytics as well as all-round business intelligence which can facilitate better decision making and execution

Quote by John Tukey



Steps For Cleaning

- There are 7 steps involved to clean and prepare the data for building predictive model.
  - Variable Identification
  - Univariate Analysis
  - Bivariate Analysis
  - Missing values treatment
  - Outlier treatment
  - Variable transformation
  - Variable creation

- The above steps could be re-iterated to prepare good data for analysis

Variable Identification

Proschool An IMS Initiative

- Understand the variables and the type of data for each variable.
- Suppose, we want to predict, the appraised value of the house for the below data. Then, we need to identify predictor variables, target variable, data type of variables and category of variables.

Name	Age	Gender	Education	Salary	Appraised Value	Location	Landacres	HouseSizes qrft	Rooms	Baths	Garage
Tony	25	M	Grad	50	700	Glen Cove	0.2297	2448	8	3.5	2
Harret	52	F	PostGrad	95	364	Glen Cove	0.2192	1942	7	2.5	1
Jane	26	F	PostGrad	65	600	Glen Cove	0.163	2073	7	3	2
Rose	45	F	Grad	100	548.4	Long Beach	0.4608	2707	8	2.5	1
John	42	M	Grad	77	405.9	Long Beach	0.2549	2042		1.5	1
Mark	62	M	PostGrad	118	374.1	Glen Cove	0.229	2089	7	2	0
Bruce	51	M	Grad	101	600	Glen Cove	0.1714	1344	8	1	0
Steve	43	M	Grad	108	299	Roslyn	0.175	1120	5	1.5	0
Carol	24	F	PostGrad	51	471	Roslyn	0.213	1817	6	2	0
Henry	25	M	PostGrad	68	510.7	Roslyn	0.1377	2496		2	1
Donald	41	M	Grad	86	517.7	Long Beach	0.2497	1615	7	2	1
Maria	51	F	Grad	122	1200	Long Beach	0.4116	4067	9	4	1
Janet	49	F	PostGrad	112	700	Roslyn	0.3372	3130	8	3	1
Sophia	32	F	Grad	85	374.8	Roslyn	0.1503	1423		2	0
Jeffery	37	M	Grad	90	543	Long Beach	0.2348	1799	6	2.5	1

Below, the variables have been defined in different category:

Type of Variable	Data Type	Variable Category

• **Predictor Variable**
  • Housesizesqrft
  • Age
  • Salary
  • Education
  • Baths
  • Room
  • Garage
• **Target Variable**
  • Appraised_value
	• **Character**
	  • Name
	  • Gender
	  • Education
	  • Location
	• **Numeric**
	  • Age
	  • Salary
	  • Baths
	  • Room
	  • Garage
	  • Landacres
	  • Housesizesqrft
	  • Appraised_value
		• **Categorical**
		  • Gender
		  • Education
		  • Location
		• **Continuous**
		  • Age
		  • Salary
		  • Landacres
		  • Housesizesqrft
		  • Appraised_value
		• **Discrete**
		  • Baths
		  • Rooms
		  • Garage

Note: Numeric variable is of two types, discrete and continuous depending on the nature of the data value that a variable takes.

Univariate Analysis

Univariate Analysis(1/2)

- Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so we analyze one variable at one time.
- It doesn’t deal with causes or relationships among variables but mostly to describe and summarize and find patterns in the data.
- Used to highlight missing and outlier values
- Method to perform univariate analysis depends on whether the variable type is categorical or continuous

Continuous Variables

These measures(below) help in determining the central value and also the dispersion of continuous variables

Central Tendency	Measure of Dispersion	Visualization Method
Mean	Range	Histogram
Median	Quartile	Box-Plot
Mode	IQR	
Min	Variance and SD	
Max	Skewness and Kurtosis	

Categorical Variables

Frequency table is used to understand the distribution of each category under a variable, we can produce count and count% against each category

Bar plots could be used to visualize the Frequency Table

Lets take a look at the univariate analysis for our dataset…

```
> data_file = read.csv("D://IMS Proschool//EDA//EDA_data.csv")
> View(data_file)
> summary(data_file)
      Name          Age           Gender    Education       Salary        AppraisedValue          Location
 Bruce   :1   Min.    :24.00   F:7      Grad    :9   Min.    : 50.00   Min.    : 299.0   Glen Cove :5
 Carol   :1   1st Qu.:29.00   M:8      PostGrad:6   1st Qu.: 72.50   1st Qu.: 390.4   Long Beach:5
 Donald  :1   Median :42.00                         Median : 90.00   Median : 517.7   Roslyn    :5
 Harret  :1   Mean   :40.33                         Mean   : 88.53   Mean   : 547.2
 Henry   :1   3rd Qu.:50.00                         3rd Qu.:104.50   3rd Qu.: 600.0
 Jane    :1   Max.   :62.00                         Max.   :122.00   Max.   :1200.0
 (Other):9
   Landacres        HouseSizesqrft       Rooms           Baths           Garage
 Min.    :0.1377   Min.    :1120   Min.    :5.000   Min.    :1.000   Min.    :0.0
 1st Qu.:0.1732   1st Qu.:1707   1st Qu.:6.750   1st Qu.:2.000   1st Qu.:0.0
 Median :0.2290   Median :2042   Median :7.000   Median :2.000   Median :1.0
 Mean   :0.2425   Mean   :2141   Mean   :7.167   Mean   :2.333   Mean   :0.8
 3rd Qu.:0.2523   3rd Qu.:2472   3rd Qu.:8.000   3rd Qu.:2.750   3rd Qu.:1.0
 Max.   :0.4608   Max.   :4067   Max.   :9.000   Max.   :4.000   Max.   :2.0
                                   NA's   :3
>
```

Missing value detected in variable Rooms and denoted as NA.

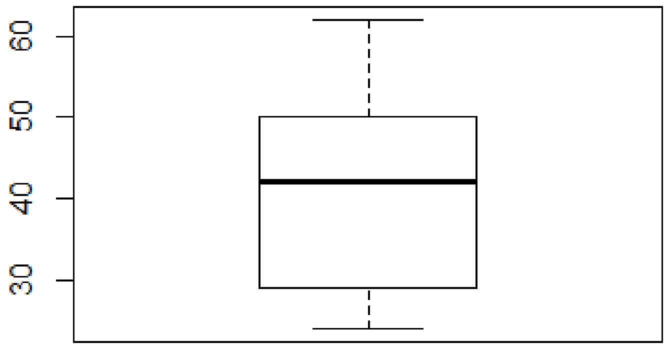
This R output gives us a tabular output which summarizes the data. We get an idea of Minimum and Maximum values along with Mean and 1st and 3rd quartile for the numeric data. We also can see if there are any missing values in any of the variables.
Similarly, for the categorical variables we get a frequency distribution.

We can view graphical output through Box Plot and Histogram for Continuous variable

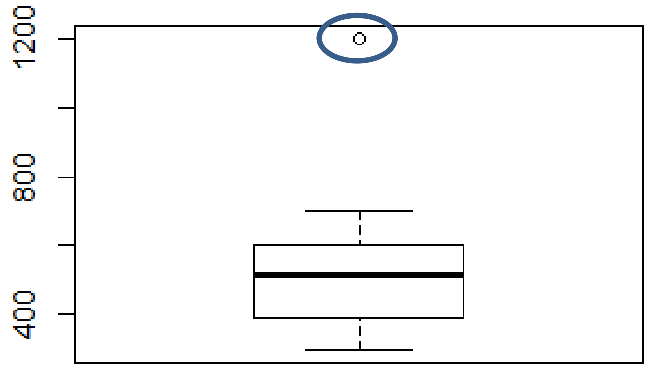
For graphical output of continuous variable

```
> boxplot(data_file$AppraisedValue)
> boxplot(data_file$Age)
> hist(data_file$Salary)
```

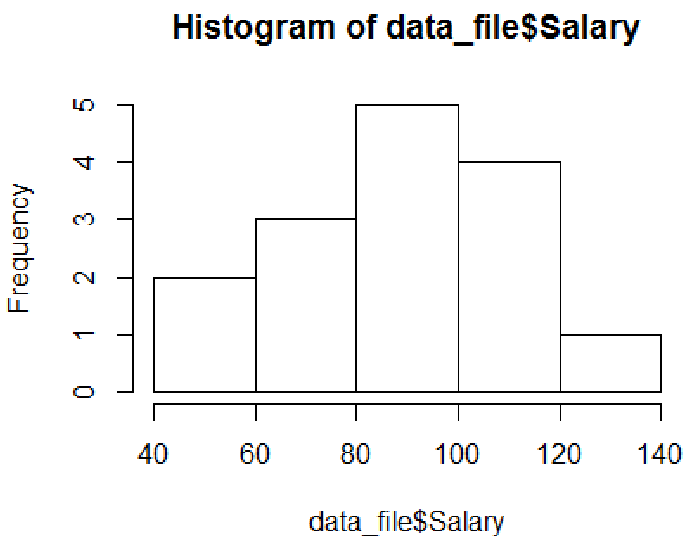
Box Plot Of variable Age

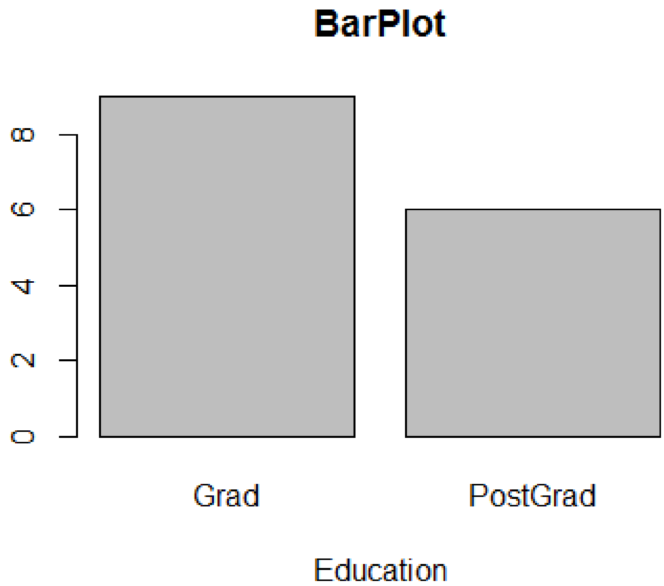


Box Plot Of variable AppraisedValue



It can be seen that Box Plot for Age has no outlier but for the variable AppraisedValue we can see one data point beyond the whiskers of the box plot, which can be denoted as a outlier. Hence, this visual representation can be used to detect outliers





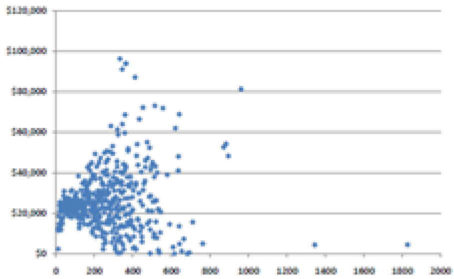
Through the Histogram, we get an idea of the distribution of the data, whether it is skewed or normally distributed. We observe that the variable Salary is roughly normally distributed. Similarly, we can plot histogram for the remaining continuous variables.

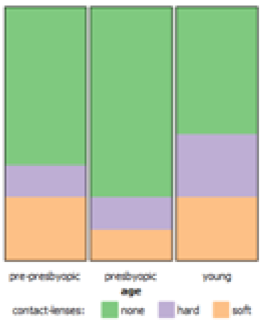
Visualization of the Categorical variable is shown through the BarPlot.
As we can see the Barplot above, we observe that our data has more number of graduates than postgraduates.

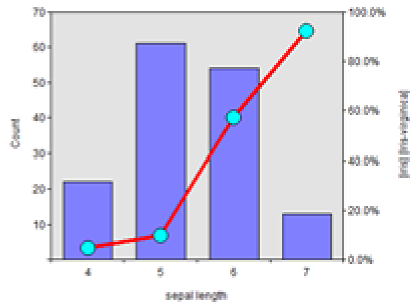
Bivariate Analysis

Bivariate Analysis

- In Univariate Analysis, we study one variable at a time, like we did in earlier slides, but if we want to find if there is any relation between two variables we need to perform bivariate analysis.
- Bivariate analysis, can be performed for any combination of categorical and continuous variables.
- Different methods are used to tackle different combinations during analysis process.
- Possible Combinations are:-
  - Continuous & Continuous
  - Continuous & Categorical
  - Categorical & Categorical

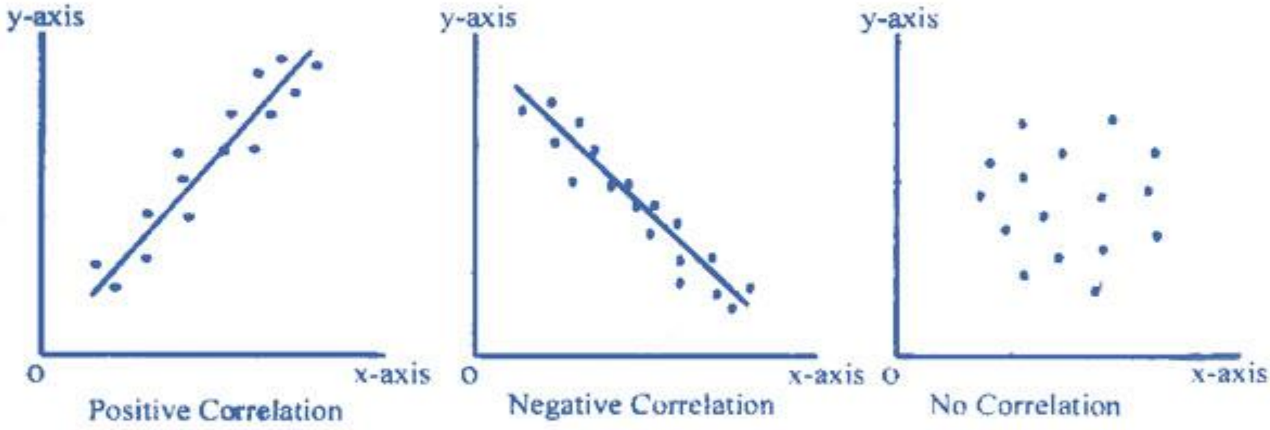






Bivariate Analysis - Continuous & Continuous(1/2)

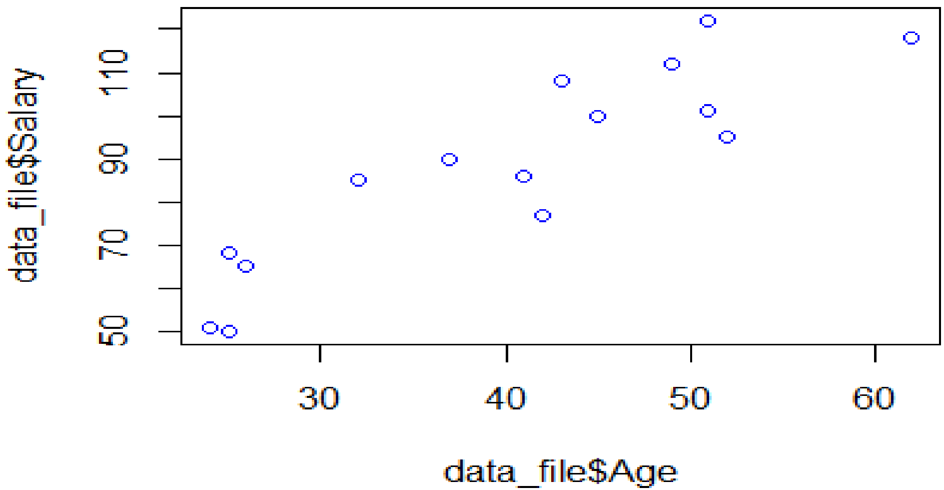
- **Scatter plot**
  - find out the relationship between two variables
  - The pattern of scatter plot indicates the relationship between variables, but does not indicates the strength of relationship amongst them
  - The relationship can be linear or non-linear
  - To find the strength of the relationship, we use Correlation(-1 negative linear correlation to +1 positive linear correlation and 0 is no correlation).
  - We get an idea of some relation and pattern among 2 variables in the dataset.



Bivariate Analysis - Continuous & Continuous(2/2)


```
#Scatter Plot of Age vs Salary

> plot(data_file$Age, data_file$Salary, col = "blue")
```



This scatter plot tells us that there is positive linear relationship between Age and Salary, these two are continuous variables.

Bivariate Analysis - Categorical & Categorical(1/3)

Methods to identify the relationship between two categorical variables.

- **Two-way table:** In this method by creating a two-way table of count and count%. Both row and column represents category of their respected variable.

- **Stacked Column Chart:** This method is one of the most visual form of Two-way table.

- **Chi-Square Test:** It derives the statistical significance of relationship between the variables for a larger population as well. The difference between the expected and observed frequencies in one or more categories in the two-way table.

#Two-way Table

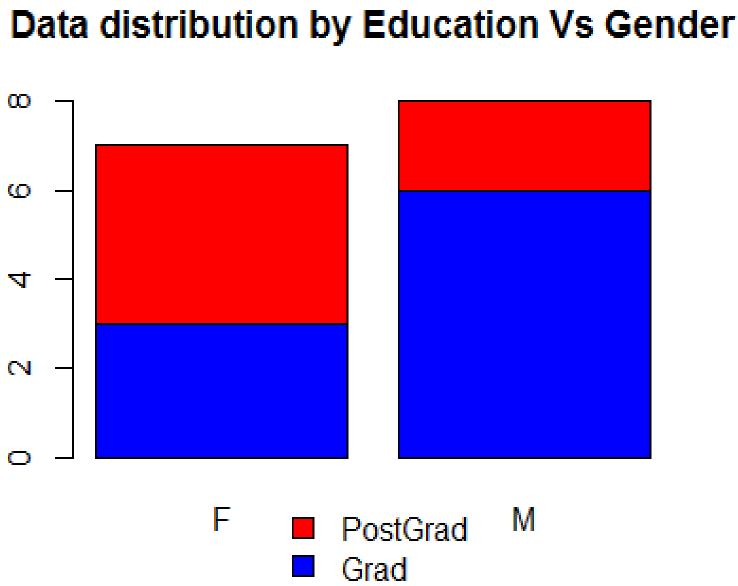
```
> counts = table(data_file$Education,data_file$Gender)
>counts
          F M
 Grad     3 6
 PostGrad 4 2
```

Through the two-way table for Education and Gender, we see that more Males are Grads whereas more Females are Post Grads.

#Stacked bar-chart

```
>barplot(counts, main = "Data distribution by Education Vs Gender",col = c("blue","red"),legend=rownames(counts),
+ args.legend = list(x = "bottom", bty = "n", inset=c(-0.40, -.40)))
```

For the two way table above we can see a stacked bar-chart.



Bivariate Analysis - Categorical & Categorical(3/3)

- Chi square test

$$X^2 = \sum (O - E)^2 / E$$

O = observed frequency
E = expected frequency

chi-square test is found by

$$E = \frac{row\ total \times column\ total}{sample\ size}$$

- If $p<0.05$ then it indicates that the relationship between the variables is significant at 95% confidence

Example

It is used to determine whether there is a significant association between the two categorical variables.

H_0: Variable Education and Variable Gender are independent.

H_a: Variable Education and Variable Gender are not independent.

Observed			
	F	M	Total
Grad	3	6	9
Post Grad	4	2	6
Total	7	8	15

Expected			
	F	M	Total
Grad	4.2	4.8	9
Post Grad	2.8	3.2	6
Total	7	8	15

(Obs-Exp)^2 / Exp	F	M
Grad	0.342857	0.3
Post Grad	0.514286	0.45

Adding up all the values from the above table, we get a chi sqr value.

Chi sqr = 0.342857 + 0.3 + 0.514286 + 0.45 = 1.607143

P-value corresponding to the above chi sqr value with 1 df and alpha = 0.05 is 0.2049.

Since p-value > 0.05, we do not reject Null hypothesis and conclude that Education and Gender are independent variables.

Missing Values

Missing Value Treatment

- There may be situations where there could be missing values in your data.

- Missing Data will not make any impact on the result if its percentage is less 1%, if missing data's range within the range of 1-5% then it is somehow manageable; however in case of 5-15% complex techniques are used for handling the problems of missing data but if it exceeds from 15% then it will surely hinder the result achieved after applying data mining techniques

- Handling such values is very important as this could lead to wrong results.

- Missing values could occur due to several reasons like,
  - During data extraction i.e. while fetching the data required for the analysis
  - During data collection itself there could be some fields for which the values may not have been collected.

- But there are ways to handle these problems

Treating Missing Values

- **Deletion**: Deleting observations or variables.

  If a particular variable is having more missing values than rest of the variables in the dataset, then we are better off without that variable unless it is a really important predictor that makes a lot of business sense.

  Also, if in a huge dataset we have very minute number of observations missing, then we can delete the whole of observations altogether.

We can delete the variable altogether since majority values are missing(NA) for it

Obs	Age	Salary (in 1000s)	Location
1	24	15	North
2	28	20	NA
3	36	45	NA
4	30	35	NA
5	25	20	South
6	35	54	NA
7	41	60	NA
8	38	52	NA
9	28	26	NA
10	29	25	NA

We can delete obs 4 and 7 from the dataset as they are very few missing(NA) values in a large dataset

Obs	Age	Salary (in 1000s)
1	24	15
2	28	20
3	36	45
4	30	NA
5	25	20
6	35	54
7	41	NA
…………		
1000	24	18

- **Single Imputation:** In single imputation, we use mean, median or mode.

  If the variable is continuous then replace the missing values with either mean, median or mode.

  If the variable is otherwise generally normally distributed (and in particular does not have any skewness), we would choose mean.

  If the data skewed, median imputation is suggested.

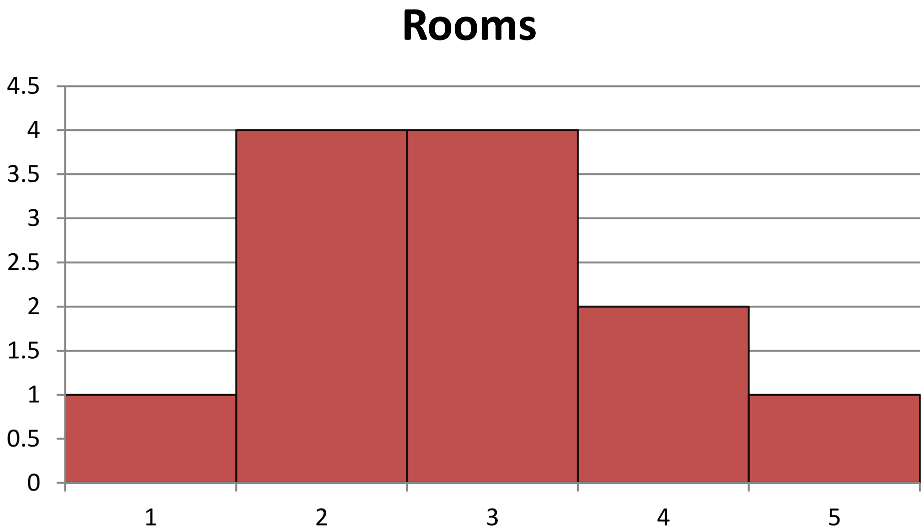
  If the variable is categorical then we could replace the missing values with the most frequent occurring value in that variable, i.e the mode.

Single Imputation - by Mean/Median/Mode

Name	Age	Gender	Education	Salary	AppraisedValue	Location	Landacres	HouseSizesqrft	Rooms	Baths	Garage
Tony	25	M	Grad	50	700	Glen Cove	0.2297	2448	8	3.5	2
Harret	52	F	PostGrad	95	364	Glen Cove	0.2192	1942	7	2.5	1
Jane	26	F	PostGrad	65	600	Glen Cove	0.163	2073	7	3	2
Rose	45	F	Grad	100	548.4	Long Beach	0.4608	2707	8	2.5	1
John	42	M	Grad	77	405.9	Long Beach	0.2549	2042		1.5	1
Mark	62	M	PostGrad	118	374.1	Glen Cove	0.229	2089	7	2	0
Bruce	51	M	Grad	101	600	Glen Cove	0.1714	1344	8	1	0
Steve	43	M	Grad	108	299	Roslyn	0.175	1120	5	1.5	0
Carol	24	F	PostGrad	51	471	Roslyn	0.213	1817	6	2	0
Henry	25	M	PostGrad	68	510.7	Roslyn	0.1377	2496		2	1
Donald	41	M	Grad	86	517.7	Roslyn	0.2497	1615	7	2	1
Maria	51	F	Grad	122	1200	Roslyn	0.4116	4067	9	4	1
Janet	49	F	PostGrad	112	700	Roslyn	0.3372	3130	8	3	1
Sophia	32	F	Grad	85	374.8	Roslyn	0.1503	1423		2	0
Jeffery	37	M	Grad	90	543	Roslyn	0.2348	1799	6	2.5	1

We can see that the variable Rooms has 3 missing values, we need to find a way to replace the missing values

Missing Values



Looking at the histogram of the variable Rooms (non missing value, we see that it is normally distributed. Hence we can impute missing values with Mean of non-missing data

```
>data_file$Rooms[is.na(data_file$Rooms)] <- mean(data_file$Rooms, na.rm = TRUE)
> View(data_file)
```

Name	Age	Gender	Education	Salary	AppraisedValue	Location	Landacres	HouseSizesqrft	Rooms	Baths	Garage
Tony	25	M	Grad	50	700.0	Glen Cove	0.2297	2448	8.000000	3.5	2
Harret	52	F	PostGrad	95	364.0	Glen Cove	0.2192	1942	7.000000	2.5	1
Jane	26	F	PostGrad	65	600.0	Glen Cove	0.1630	2073	7.000000	3.0	2
Rose	45	F	Grad	100	548.4	Long Beach	0.4608	2707	8.000000	2.5	1
John	42	M	Grad	77	405.9	Long Beach	0.2549	2042	**7.166667**	1.5	1
Mark	62	M	PostGrad	118	374.1	Glen Cove	0.2290	2089	7.000000	2.0	0
Bruce	51	M	Grad	101	600.0	Glen Cove	0.1714	1344	8.000000	1.0	0
Steve	43	M	Grad	108	299.0	Roslyn	0.1750	1120	5.000000	1.5	0
Carol	24	F	PostGrad	51	471.0	Roslyn	0.2130	1817	6.000000	2.0	0
Henry	25	M	PostGrad	68	510.7	Roslyn	0.1377	2496	**7.166667**	2.0	1
Donald	41	M	Grad	86	517.7	Long Beach	0.2497	1615	7.000000	2.0	1
Maria	51	F	Grad	122	1200.0	Long Beach	0.4116	4067	9.000000	4.0	1
Janet	49	F	PostGrad	112	700.0	Roslyn	0.3372	3130	8.000000	3.0	1
Sophia	32	F	Grad	85	374.8	Roslyn	0.1503	1423	**7.166667**	2.0	0
Jeffery	37	M	Grad	90	543.0	Long Beach	0.2348	1799	6.000000	2.5	1

```
#Other Method to impute using Hmisc Package
> library(Hmisc)
> impute(data_file$Rooms, mean) # replace with mean
[1] 8.000000 7.000000 7.000000 8.000000 7.166667 7.000000 8.000000 5.000000 6.000000
7.166667 7.000000
[12] 9.000000 8.000000 7.166667 6.000000
```

Treating Missing Values

Prediction Imputation:
Regression - In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. Next, we create a model to predict target variable based on other attributes of the non-missing data set and populate missing values of other data set
KNN Imputation- For every observation to be imputed, it identifies 'k' closest observations based on the euclidean distance and computes the weighted average of these 'k' obs.
(Note: We will get a clear idea of these methods in the course further)

Constant: This choice allows us to provide our own default value to fill in the gaps. This might be an integer or real number for numeric variables, or else a special marker or the choice of something other than the majority category for Categorical variables.

Closest fit: The closet fit algorithm depends upon exchanging absent values with present value of the similar attribute of other likewise cases. Main notion is to find out from dataset likewise scenarios and select the likewise case to the case in discussion with missing attribute values.

Treating Missing Values :: Closest Fit

Area Sq. ft	Rent
275	8000
500	10000
850	12000
900	
1000	17000
1225	19000
1500	20000
Missing value is for 900.	
The value closer to 900 with a non missing rent value is 800	
So we replace the missing rent value with 12000	
Area Sq. ft	**Rent**
275	8000
500	10000
850	12000
900	12000
1000	17000
1225	19000
1500	20000

Note:
This method is more useful for a small dataset

Business Analytics

Outliers

Proschool

An IMS Initiative

Outliers (1/2)

- What is an Outlier?

  Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

- Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set:

  - It increases the error variance and reduces the power of statistical tests
  - If the outliers are non-randomly distributed, they can decrease normality
  - They can bias or influence estimates that may be of substantive interest

Outliers (2/2)

Causes of outliers

- **Data Entry Errors -** Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.
- **Measurement Error -** When the measurement instrument used turns out to be faulty.
- **Intentional Error -** This is commonly found in self-reported measures that involves sensitive data.
- **Data Processing Error -** When data is collected from different sources
- **Sampling Error -** Data considered which is not part of the sample
- **Natural Outlier -** When an outlier is not artificial (due to error), it is a natural outlier.

Example

Let's examine what can happen to a data set with outliers. For the sample data set:
1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4
We find the following mean, median, mode, and standard deviation:
Mean = 2.58
Median = 2.5
Mode = 2
Standard Deviation = 1.08

If we add an outlier to the data set:
1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 400
The new values of our statistics are:
Mean = 35.38
Median = 2.5
Mode = 2
Standard Deviation = 114.74

As we can see, having outliers often has a significant effect on your mean and standard deviation. Because of this, we must take steps to remove outliers from our data sets.

Example

Suppose you want to take admission in a MBA school and your criteria for selection of the best MBA school is the average package received by the students.

School1
Student size: 20
Packages(in lakhs p.a.): 10,9,7,10,5,5,9,9,8,5,8,9,7,9,9,10,8,5,8,10
Avg. Package = 8

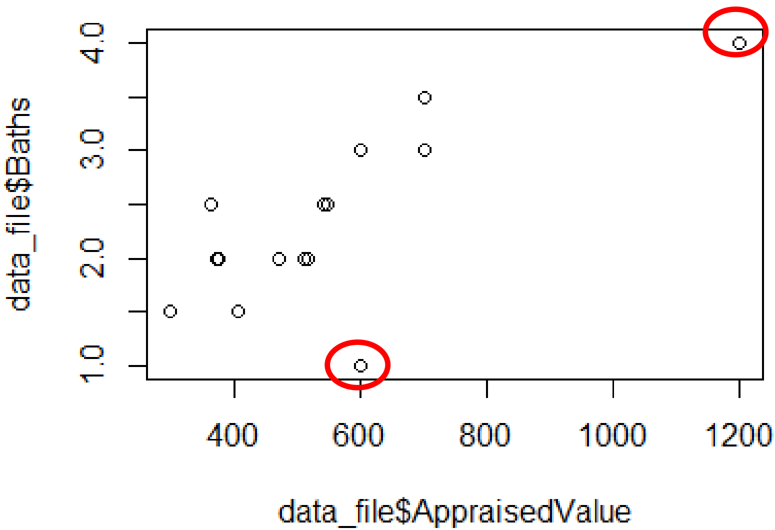
School2
Student size: 20
Packages(in lakhs p.a.): 7,6,8,10,10,10,9,**50**,9,7,50,8,7,10,7,8,8,**50**,6,8
Avg.Package = 12.4

Looking at the numbers we would decide that School 2 is the best, but the average package of school 2 has gone up just because two students got hired by an MNC(say Google).
These are outlier which are skewing our average on the higher side.

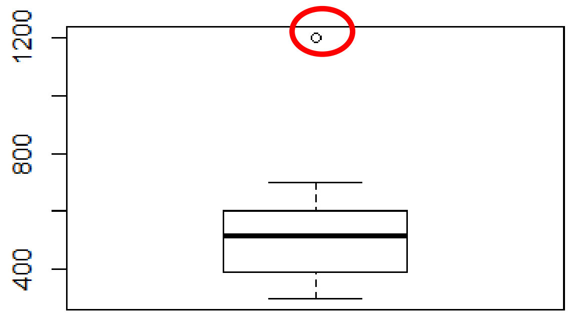
Outlier Detection - Viz

- Outliers can be detected using boxplots and scatter plots
- In our data, we plot a scatter plot for Appraised_value and Baths(bivariate analysis) and also a boxplot for Appraised_value(Univariate analysis)

Scatter plot of Appraised_value and Baths

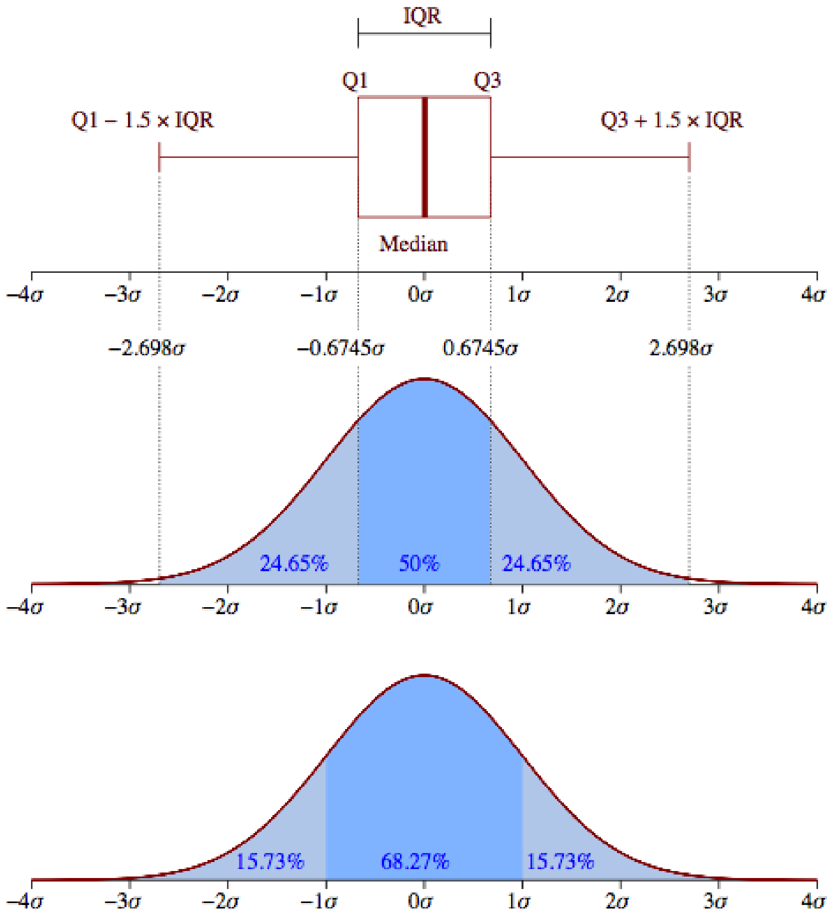


Box plot of Appraised_value



Outlier Detection – Thumb Rules

- Other than the plots, Outliers can also be detected by using certain thumb rules,
  - Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR where IQR = Q3-Q1
  - Any value which out of range of 5th and 95th percentile can be considered as outlier
  - Data points, three or more standard deviation away from mean are considered outlier.



For our data, the we find the IQR

```
> data_file = read.csv("D://IMS Proschool//EDA//EDA_data.csv")
> summary(data_file$AppraisedValue)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  299.0   390.4   517.7   547.2   600.0  1200.0
>
```

Q3 = 600, Q1 = 390

IQR = Q3-Q1 = 600-390 = 290.

Outliers will be values lying above +1.5*IQR or below -1.5*IQR.

-1.5*IQR = -1.5*290 = -435

1.5*IQR = 1.5*290 = 435. So any value lying outside the range of (-435,435) will be an outlier. Boxplot considers IQR approach for detecting Outliers.

In our data Max value for AppraisedValue var is 1200, which will be considered as an outlier!!

Looking at the 5th and 95th percentile approach for detection of Outliers :

```
>quantile(data_file$AppraisedValue, .95)
```

95%

850

```
>quantile(data_file$AppraisedValue, .05)
```

5%

344.5

Values lying below 344.5 and above 850 will be considered as outliers according to this approach.

Handle Outliers

- We could remove the outliers from the data if they are due to data entry or data processing errors
- Based on business understanding you could also replace the outliers with mean or median
- If there is a pattern of interest in the outliers then they could be handled separately. For example if the outliers are like in groups then treat both groups as two different groups and build individual model for both groups and then combine the output.
- Also the outliers can be capped with 5^{th} or 95^{th} percentile.

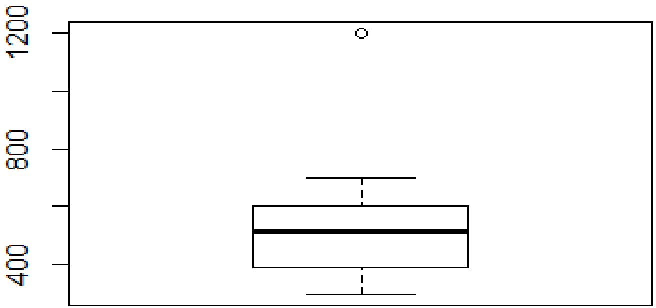
We will treat outlier in our data using R.

We will be using capping method for imputation of the outlier in our data for variable Apparaised_value

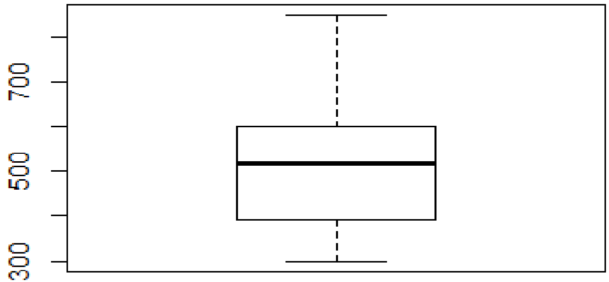
```
> boxplot(data_file$AppraisedValue)
> quantile(data_file$AppraisedValue, .95)  # 95th percentile
95% 850
> summary(data_file$AppraisedValue)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
299.0 390.4 517.7 547.2 600.0 1200.0

 #Capping outliers with the 95th percentile
> data_file$Age = ifelse(data_file$AppraisedValue >= 1000,850,data_file$AppraisedValue)
> boxplot(final_data$Age)
```

Box-Plot : Before capping



Box-Plot : After capping



Business Analytics

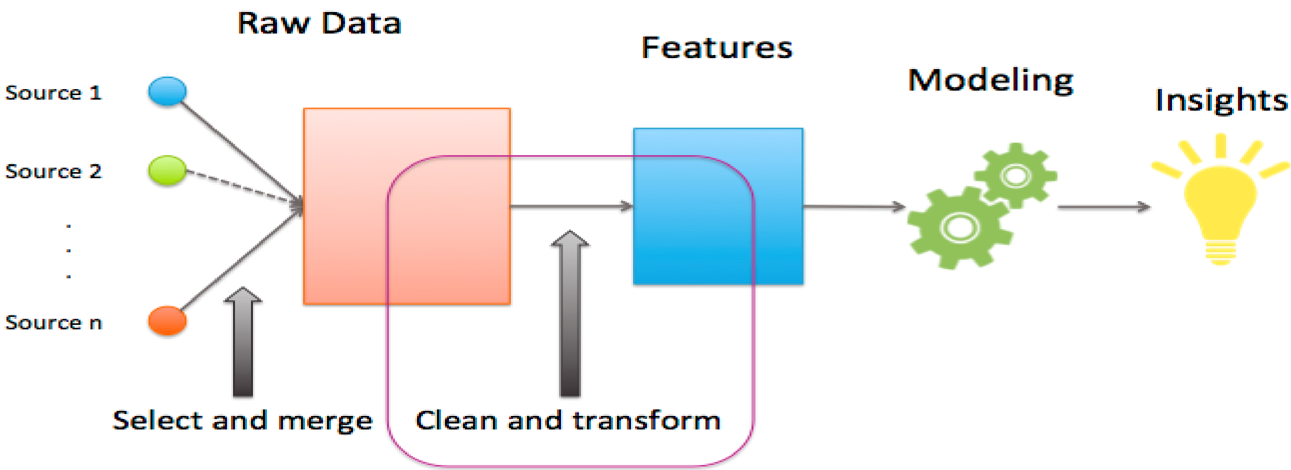
Feature Engineering

Proschool

An IMS Initiative

Feature Engineering

- Feature engineering is the science (and art) of extracting more information from existing data.
- Example
  - Several variables could be generated from a date variable i.e. Day, month, year, day of the week etc. This information helps a lot in getting idea about different characteristics of the data under study
- It can be divided into two steps,
  - Variable Transformation
  - Variable Creation



Feature Engineering – Variable Transformation

- In data modelling, transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation.

- When do we transform?
  - When we want to change the scale of a variable or standardize the values of a variable for better understanding. While this transformation is a must if you have data in different scales
  - This transformation does not change the shape of the variable distribution
  - Existence of a linear relationship between variables is easier to comprehend compared to a non-linear or curved relation.
  - Variables can be transformed by applying functions like log, square, cube etc. These transformations help in reducing skewness. For right skewed distribution, we take square / cube root or logarithm of variable and for left skewed, we take square.

Feature Engineering - Variable and Dummy Variable Creation

- Variable creation is a process to generate a new variables / features based on existing variable(s)
- Dummy coding provides one way of using categorical predictor variables in various kinds of estimation models (see also effect coding), such as, linear regression. Dummy coding uses only ones and zeros to convey all of the necessary information on group membership.
- Below is an example of variable creations (Yellow columns are original variables and the columns in blue are variables created from them)

ID	Gender	Date	Day	Month	Year	Dummy_Male	Dummy_Female
1	Male	10 May 2016	10	5	2016	1	0
2	Female	15 July 2016	15	7	2016	0	1
3	Male	01 June 2016	1	6	2016	1	0
4	Male	04 January 2016	4	1	2016	1	0
5	Female	27 March 2016	27	3	2016	0	1

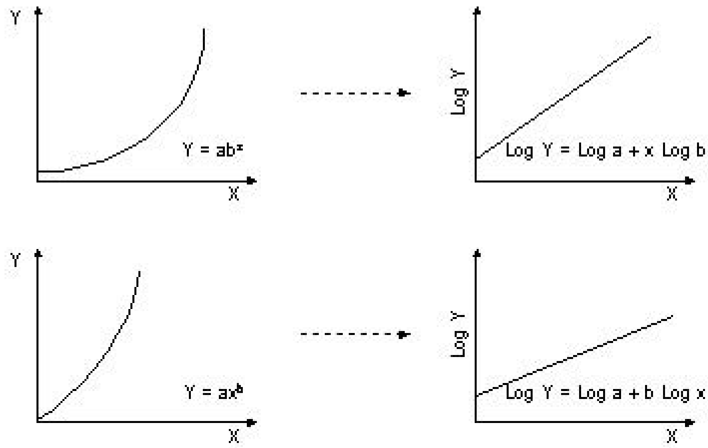
Since, Location is a categorical variable, we need to convert it to a dummy variable so that we will be able to use it as a predictor

```
> new = dummy(data_file$Location)
> View(new)
> new_data = cbind(data_file,new)
> View(new_data)
```

Name	Age	Gender	Education	Salary	AppraisedValue	Location	Landacres	HouseSizesqrft	Rooms	Baths	Garage	Location GlenCove	Location LongBeach	Location Roslyn
Tony	25	M	Grad	50	700.0	Glen Cove	0.2297	2448	8.000000	3.5	2	1	0	0
Harret	52	F	PostGrad	95	364.0	Glen Cove	0.2192	1942	7.000000	2.5	1	1	0	0
Jane	26	F	PostGrad	65	600.0	Glen Cove	0.1630	2073	7.000000	3.0	2	1	0	0
Rose	45	F	Grad	100	548.4	Long Beach	0.4608	2707	8.000000	2.5	1	0	1	0
John	42	M	Grad	77	405.9	Long Beach	0.2549	2042	7.166667	1.5	1	0	1	0
Mark	62	M	PostGrad	118	374.1	Glen Cove	0.2290	2089	7.000000	2.0	0	1	0	0
Bruce	51	M	Grad	101	600.0	Glen Cove	0.1714	1344	8.000000	1.0	0	1	0	0
Steve	43	M	Grad	108	299.0	Roslyn	0.1750	1120	5.000000	1.5	0	0	0	1
Carol	24	F	PostGrad	51	471.0	Roslyn	0.2130	1817	6.000000	2.0	0	0	0	1
Henry	25	M	PostGrad	68	510.7	Roslyn	0.1377	2496	7.166667	2.0	1	0	0	1
Donald	41	M	Grad	86	517.7	Long Beach	0.2497	1615	7.000000	2.0	1	0	1	0
Maria	51	F	Grad	122	1200.0	Long Beach	0.4116	4067	9.000000	4.0	1	0	1	0
Janet	49	F	PostGrad	112	700.0	Roslyn	0.3372	3130	8.000000	3.0	1	0	0	1
Sophia	32	F	Grad	85	374.8	Roslyn	0.1503	1423	7.166667	2.0	0	0	0	1
Jeffery	37	M	Grad	90	543.0	Long Beach	0.2348	1799	6.000000	2.5	1	0	1	0

Feature Engineering - Variable Transformation

- If the response variable is **not a linear function of the predictors**, try a different function. For example, polynomial regression involves transforming one or more predictor variables while remaining within the multiple linear regression framework.
- For another example, applying a logarithmic transformation to the response variable also allows for a nonlinear relationship between the response and the predictors while remaining within the multiple linear regression framework.
- Transforming response and/or predictor variables therefore has the potential to remedy a number of model problems
- The use of transformation will be more clear in the further course when we deal with model building.



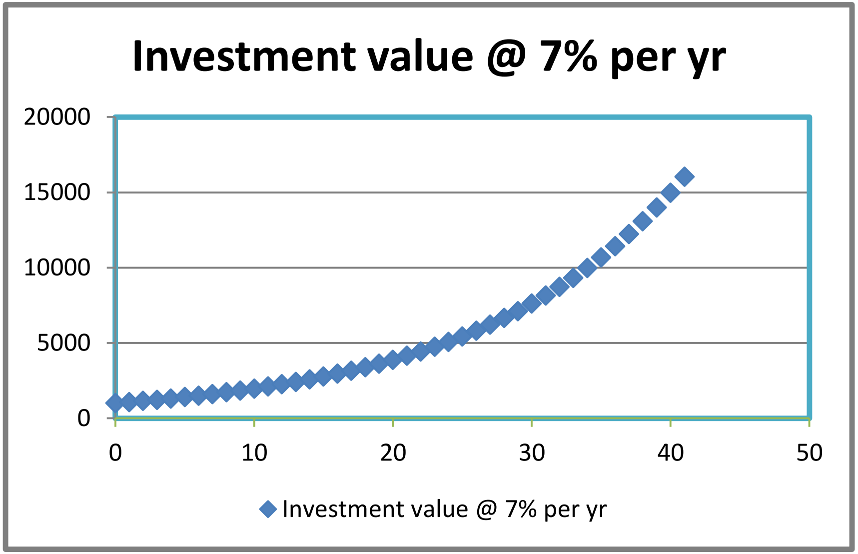
Transforming a variable involves using a mathematical operation to change its measurement scale.
In regression, a transformation to achieve linearity is a special kind of nonlinear transformation. It is a nonlinear transformation that *increases* the linear relationship between two variables.

Methods of Transforming Variables to Achieve Linearity:
There are many ways to transform variables to achieve linearity for regression analysis. Some common methods are summarized below.

Method	Transformation(s)	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	Dependent variable = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	Dependent variable = sqrt(y)	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	Dependent variable = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	Independent variable = log(x)	$y= b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	Dependent variable = log(y) Independent variable = log(x)	$\log(y)= b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

Example

Year	Investment value @ 7% per yr	Year	Investment value @ 7% per yr
0	1000	21	4141
1	1070	22	4430
2	1145	23	4741
3	1225	24	5072
4	1311	25	5427
5	1403	26	5807
6	1501	27	6214
7	1606	28	6649
8	1718	29	7114
9	1838	30	7612
10	1967	31	8145
11	2105	32	8715
12	2252	33	9325
13	2410	34	9978
14	2579	35	10677
15	2759	36	11424
16	2952	37	12224
17	3159	38	13079
18	3380	39	13995
19	3617	40	14974
20	3870	41	16023

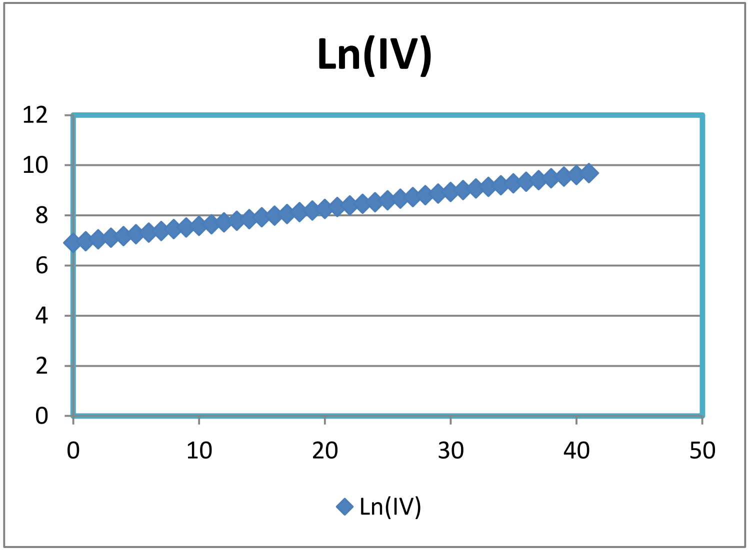


We observe that Investment value is increasing exponentially.
So, var X and Y are exponentially related.

To make them linearly dependent on each other, we need to transform the variable(Investment Value)

After Transformation with LN

Year	Ln(Investment value)	Year	Ln(Investment value)
0	6.907755279	21	8.328692584
1	6.975413927	22	8.396154863
2	7.043159916	23	8.464003363
3	7.110696123	24	8.531490496
4	7.178545484	25	8.599141774
5	7.24636808	26	8.666819365
6	7.313886832	27	8.73456009
7	7.381501895	28	8.802221746
8	7.448916103	29	8.869819953
9	7.516433303	30	8.937481228
10	7.584264818	31	9.005159521
11	7.652070746	32	9.072800958
12	7.719573989	33	9.140454245
13	7.787382026	34	9.208137948
14	7.855157006	35	9.275847174
15	7.922623574	36	9.343471685
16	7.990238186	37	9.411156511
17	8.058010801	38	9.478763169
18	8.125630988	39	9.546455402
19	8.193400232	40	9.614070643
20	8.261009786	41	9.681780469



We have transformed Investment value by LN

After transformation the vars are linearly related to each other.

We will learn further on data transformation in further course.

Thank You