Business Analytics

Exploratory Data Analysis & Data Cleaning

Proschool
An IMS Initiative

Why Clean Data?

- Data exploration helps to improve model's accuracy

- Spending significant time on exploration and analyzing data is important.

  Garbage in → Garbage Out

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

Steps For Cleaning

- There are 7 steps involved to clean and prepare the data for building predictive model.
  - Variable Identification
  - Univariate Analysis
  - Bi-variate 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

- Understand the variables and the type of data for each variable

- Identify Predictor(Input) and Target(output)

Univariate Analysis

Proschool
An IMS Initiative

Univariate Analysis(1/2)

- Exploring variables one by one.
- Used to highlight missing and outlier values
- Method to perform univariate analysis depends on whether the variable type is categorical or continuous
- **Continuous Variables**
  - The measures help in determining the central value and also the dispersion of the data.

Measures	Visualization Method
Mean	Histogram
Median	Box Plot
Mode	
Min	
Max	
Range	
Quartile	
IQR	
Variance	
Standard deviation	

Univariate Analysis(2/2)

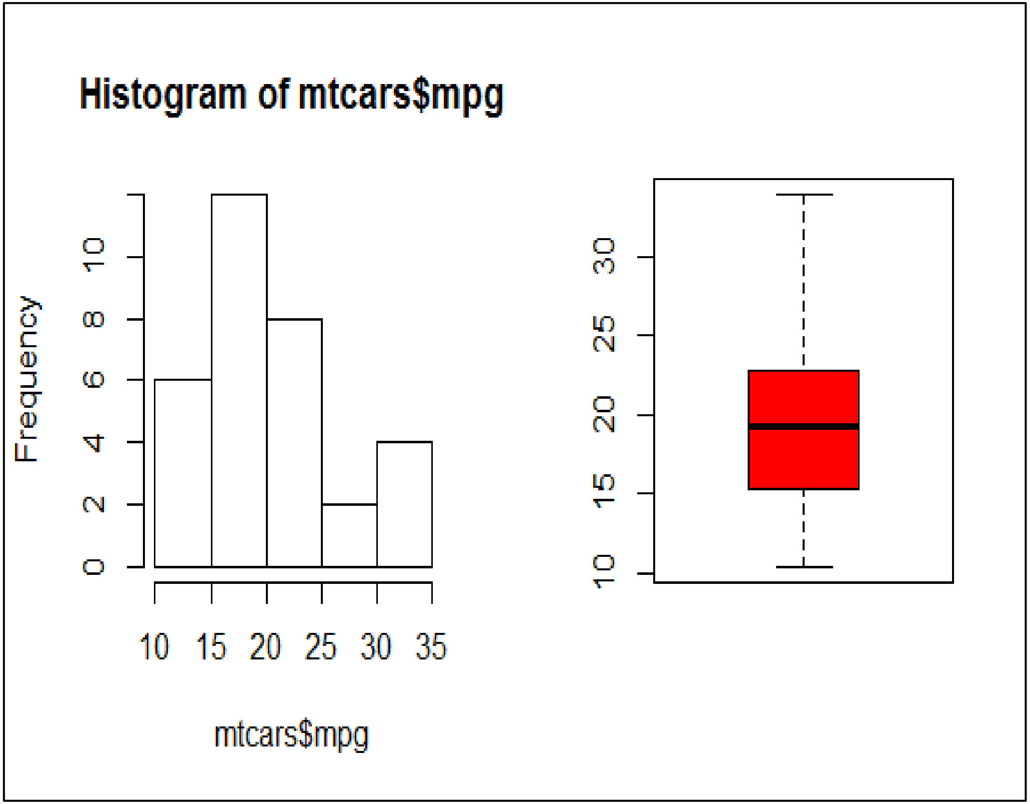
- Categorical Variables
  - frequency table is used to understand the distribution of each category
  - Bar plots could be used to visualize the counts
  - Measured in two metrics, Count and Count% against each category

Univariate Analysis in R


```
> data(mtcars)
> summary(mtcars)
      mpg             cyl             disp             hp             drat             wt             qsec
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325   Median :17.71
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90
       vs               am              gear            carb
 Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
 Median :0.0000   Median :0.0000   Median :4.000   Median :2.000
 Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
> hist(mtcars$mpg)
> boxplot(mtcars$mpg, col="red")
```

Univariate Analysis in R



Bivariate Analysis

Bivariate Analysis

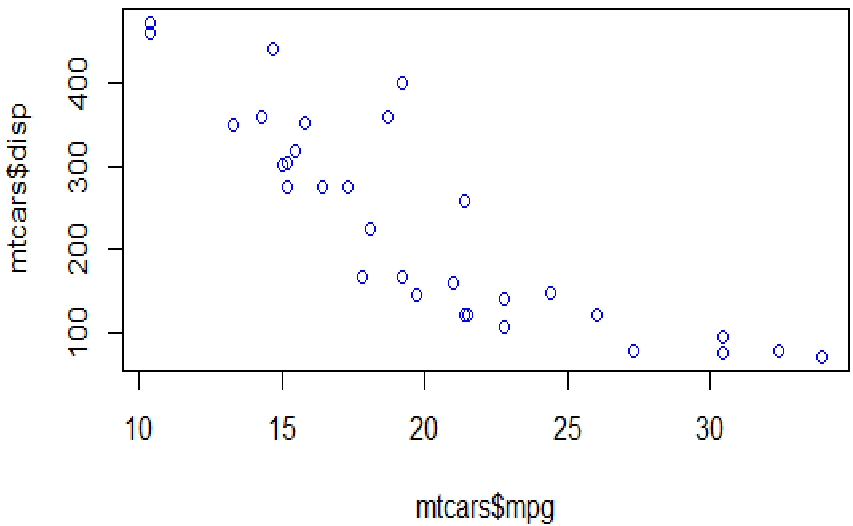
- Finds out the relationship between two variables
- Can be performed for any combination of categorical and continuous variables.
- Different methods are used to tackle different combinations during analysis process.
- Possible Combinations:-
  - Continuous & Continuous
  - Continuous & Categorical
  - 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
  - The relationship can be linear or non-linear
  - Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them
  - To find the strength of the relationship, we use Correlation(-1 negative linear correlation to +1 positive linear correlation and 0 is no correlation)
  - Correlation formula: Correlation = Covariance(X,Y) / SQRT(Var(X)* Var(Y))

Bivariate Analysis - Continuous & Continuous(2/2)

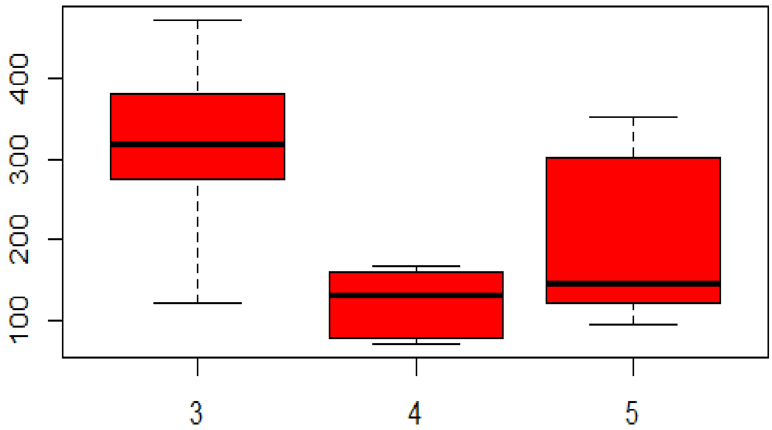
```
> plot(mtcars$mpg, mtcars$disp, col="blue")
```



Bivariate Analysis - Continuous & Categorical

- Boxplot
  - Plot the categorical variable on the x axis and the continuous variable on the y axis

```
> boxplot(disp~gear, col="red")
```



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.

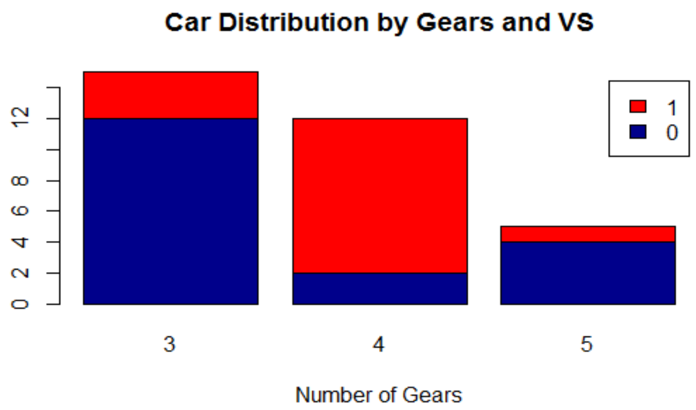
- **Chi-Square Test:** It derives the statistical significance of relationship between the variables for a larger population as well. The difference between the expected and observed frequencies in one or more categories in the two-way table.

Bivariate Analysis - Categorical & Categorical(2/3)

```
> counts = table(mtcars$vs, mtcars$gear)
> counts

     3  4  5
  0 12  2  4
  1  3 10  1
```

```
> barplot(counts, main="Car Distribution by Gears and VS",
+         xlab="Number of Gears", col=c("darkblue","red"),
+         legend = rownames(counts))
```



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

Missing Values

Missing Value Treatment

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

- 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

- If the dataset has lot of records then we could have the freedom of deleting the entire record where missing values are there

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

- If the variable is categorical then we could replace the missing values with the most frequent occurring value in that variable

Outliers

Outliers (1/2)

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

- 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.

Outliers (2/2)

- **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.

Outliers Impact

- 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.
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

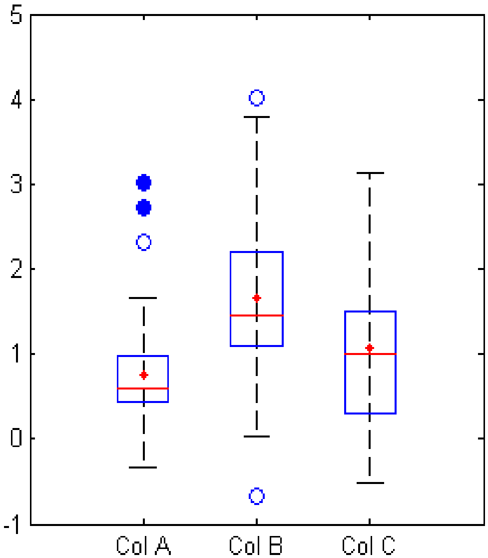
Outlier Impact Example

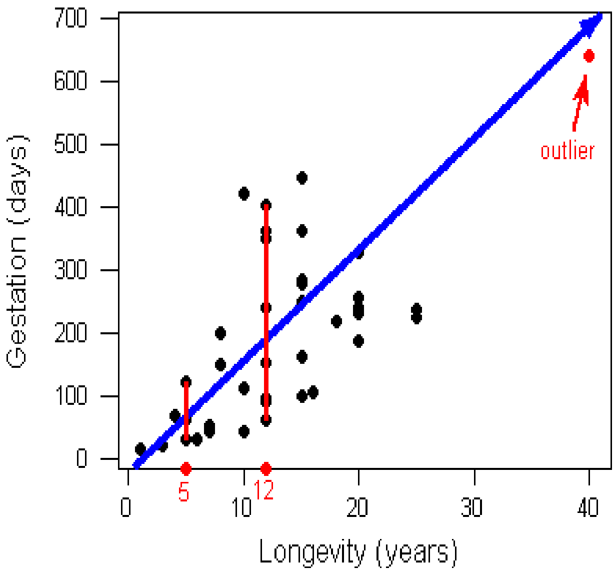
Proschool An IMS Initiative

	Without Outlier	WithOutlier
	1	100
	2	2
	3	3
	4	4
	5	5
Mean	3.00	22.80
Median	3	4
Std. Dev.	1.58	43.17

Outlier Detection - Viz

- Outliers can be detected using boxplots and scatter plots
- **EX: 1.**The average monthly income of customers is Rs.30,000. But there are also people with monthly income of Rs.5000 and Rs.5L which will be outliers.





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.

Handle Outliers

Proschool An IMS Initiative

- 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.

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 Creation

Proschool
An IMS Initiative

- Variable creation is a process to generate a new variables / features based on existing variable(s)
- 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

Thank You