**Data Preprocessing:**

summary() – To get descriptive statistics for each variable.

Sometimes numeric variables are used for other purpose which doesn’t make any sense, so we change that to factor variable using as.factor

**Handle or cleaning missing values:**

1. Handle by removing record
2. Replace the missing value with some constant, specified by the analyst
3. Replace the missing value with the field mean (for numeric variables) or the mode (for categorical variables)
4. Replace the missing values with a value generated at random from the observed distribution of the variable.
5. Replace the missing values with imputed values based on the other characteristics of the record.

**Data manipulation:**

Min-max normalization

Convert the values to be between 0 and 1

x ∗ i = xi − min(X) max(X) − min(X)

z-score standardization

Transformations to achieve normality

Log; srqt and inverse sqrt transformation

Create a function for calculating skewness

skewness = 3(mean − median) sd

skewness <- function(x) { return(3 \* (mean(x) - median(x))/sd(x)) }

skewness of -0.8 to 0.8 is acceptable for data normality

Skewness is a measure of the asymmetry of the distribution of a variable

A positive skew value indicates that the tail on the right side of the distribution is longer than the left side and the bulk of the values lie to the left of the mean

A negative skew value indicates that the tail on the left side of the distribution is longer than the right side and the bulk of the values lie to the right of the mean.

**Exploratory Data Analysis:**

Lattice vs GGplot

**Scatterplot?**

**Regression Analysis:**

1. **Linear Regression – For modeling numerical values**

Linear Regression is about quantifying the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable

**Correlation** describes the strength of the \*linear\* association between two variables.

cor(poverty$Poverty , poverty$Graduates)

* **Takes values between -1 and +1**
* A value of **0** indicates no linear association.
* correlation coefficient closest to +1 or-1 will have strong correlation
* The relationship between two variables is modeled using a straight line.
* **i.e.** y = β1 ∗ X + β0 function of Regression line
* **Residuals le**ftovers from the model fit. Difference between observed and predicted
* We want a line that has small residuals
* **After building linear model get the coefficients,**
* my\_lm$coefficient

(Intercept) Graduates 64.7809658 -0.6212167

**Slope Interpretation:**

For each additional% point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

**Intercept:**

Intercept States with no HS graduates are expected on average to have 64.68%of residents living below the poverty line.

my\_lm$coefficients (Intercept) Graduates 64.7809658 -0.6212167

**Prediction** Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction,

**Extrapolation** Applying a model estimate to values outside of the realm of the original data is called extrapolation. Sometimes the intercept might be an extrapolation

**R square: The strength of fit of the linear model is best explained by** R.square

**R square is calculated as square of correlation coefficient.**

It tells what percent of variability in the response variable is explained by the model.

**Low p value means significant results**

**Prediction function:**

prediction = function(x, my\_model) {

pr = my\_model$coefficients[1] + x \* my\_model$coefficients[2]

return(as.numeric(pr))

}

prediction(84, lm\_pov\_grad)

1. **Multiple Regression or Multivariate**

Multiple variables: y and x1, x2, ···

**Scatterplot is xyplot**

R2 = explained variability/ total variability

Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

When any variable is added to the model R2 increases. I But if the added variable doesn’t really provide any new information, or is completely unrelated, adjusted R2 does not increase.

**Model selection:**

1. **R2 adj approach:** I Start with the full model I Drop one variable at a time and record R2 adj of each smaller model I Pick the model with the highest increase in R2 adj I Repeat until none of the models yield an increase in R2 adj
2. **p-value approach:** I Start with the full model I Drop the variable with the highest p-value and refit a smaller model I Repeat until all variables left in the model are significant

**Classification Models:**

**Notes:**

R-square value tells you how much variation is explained by your model.

So 0.1 R-square means that your model explains 10% of variation within the data.

The greater R-square the better the model.

Whereas p-value tells you about the F statistic hypothesis testing of the "fit of the intercept-only model and your model are equal".

So if the p-value is less than the significance level (usually 0.05) then your model fits the data well.

Thus you have four scenarios:

1) low R-square and low p-value (p-value <= 0.05)

2) low R-square and high p-value (p-value > 0.05)

3) high R-square and low p-value

4) high R-square and high p-value

Interpretation:

1) means that your model doesn't explain much of variation of the data but it is significant (better than not having a model)

2) means that your model doesn't explain much of variation of the data and it is not significant (worst scenario)

3) means your model explains a lot of variation within the data and is significant (best scenario)

4) means that your model explains a lot of variation within the data but is not significant (model is worthless)

R squared is about explanatory power;

the p-value is the "probability" attached to the likelihood of getting your data results (or those more extreme) for the model you have

a *p*-value to weigh the strength of the evidence

*p*-value is a number between 0 and 1

* A small *p*-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
* A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
* *p*-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

**P** tells about **statistical significance**.

A **confusion matrix** is a table that is often used to describe the performance of a classification **model** (or "classifier") on a set of test data for which the true values are known

ANOVA AND SUMMARY DIFFERENCE?

**ANOVA is all about comparing two models. You can literally compare any two models in the following manner**

**Advantages of DT**

1. Easy to Understand: Decision tree output is very easy to understand, and It does not require any statistical knowledge to read and interpret them.
2. Useful in Data exploration: Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable. For example, we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.
3. Less data cleaning required: It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
4. Data type is not a constraint: It can handle both numerical and categorical variables.

Implicitly perform feature selection. – Discover nonlinear relationships and interactions.

– Require relatively little effort from users for data preparation:

• they do not need variable scaling;

• they can deal with a reasonable amount of missing values;

• they are not affected by outliers. – Easy to interpret and explain. –

Can generate rules helping experts to formalize their knowledge?

**Data with more numeric use KNN as we calculate k value ;**

**Data is more of classification use DT**

**Handle missing value:**

is.null - to detect NULL variables.

is.na – to detect Not available variables.

**NaN- Not a Number** is.nan

na.omit(DT, cols="x")

na.omit(DT, cols=c("x", "y"))

**scatter plot**

plot(mlb11$runs ~ mlb11$at\_bats, main = "Relationship between Runs and atBats", xlab = "At Bats", ylab = "Runs")

Regression Coefficients: Typically, the coefficient of a variable is interpreted as the change in the response based on a 1-unit change in the corresponding explanatory variable keeping all other variables held constant.