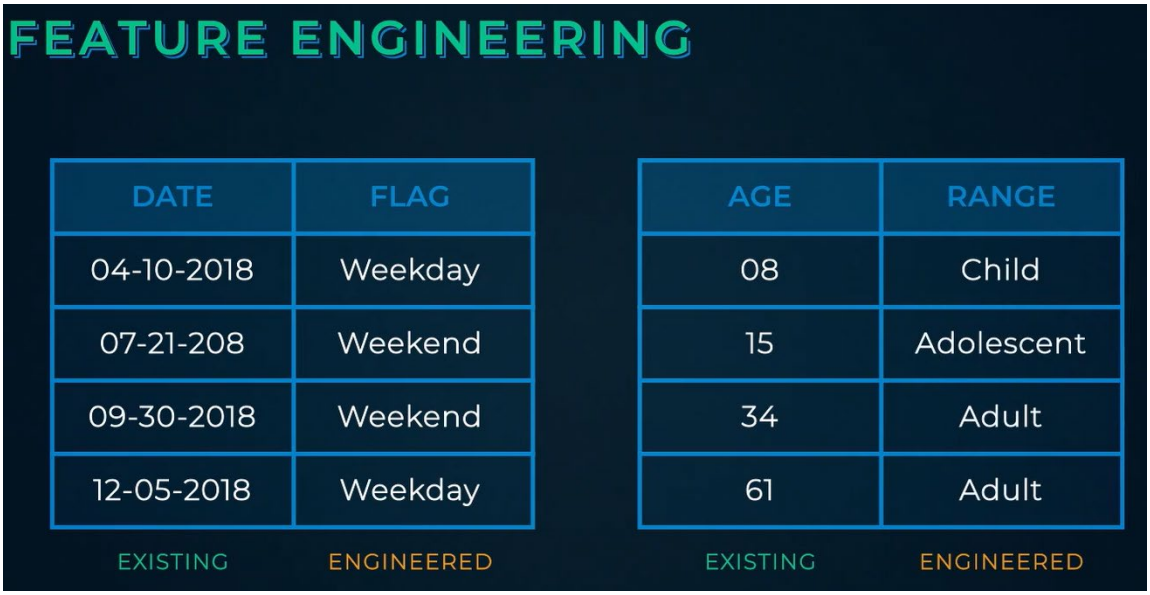
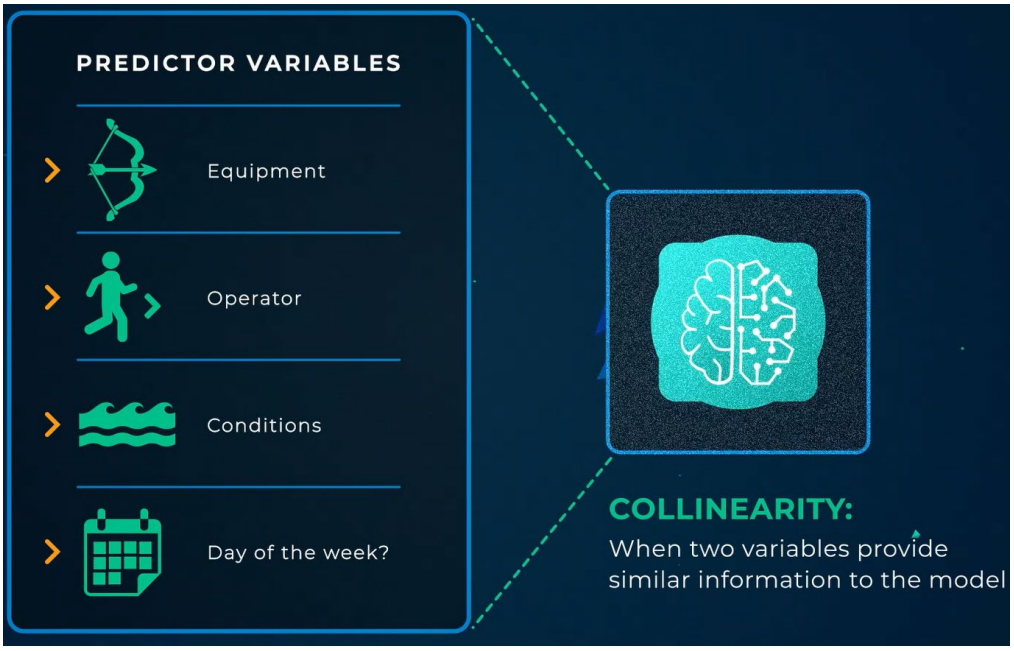
Predictive Master Notes

Data Investigation Concepts

Feature Engineering is using the data you have to generate the data you don't.



Collinearity is when two variables provide similar information to the model.

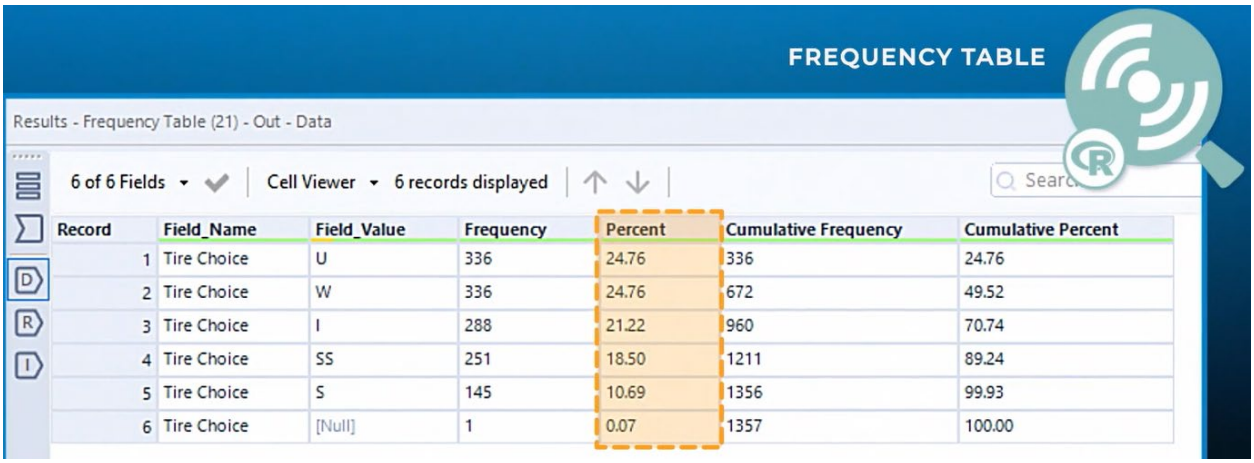


The simplest model is preferred when performance between models is similar (i.e., parsimony). If your historical data has a field that won't be captured going forward, then you don't want to include that field in the model (this is a means of "future proofing").

The *Field Summary* tool is a great tool for investigating a dataset.

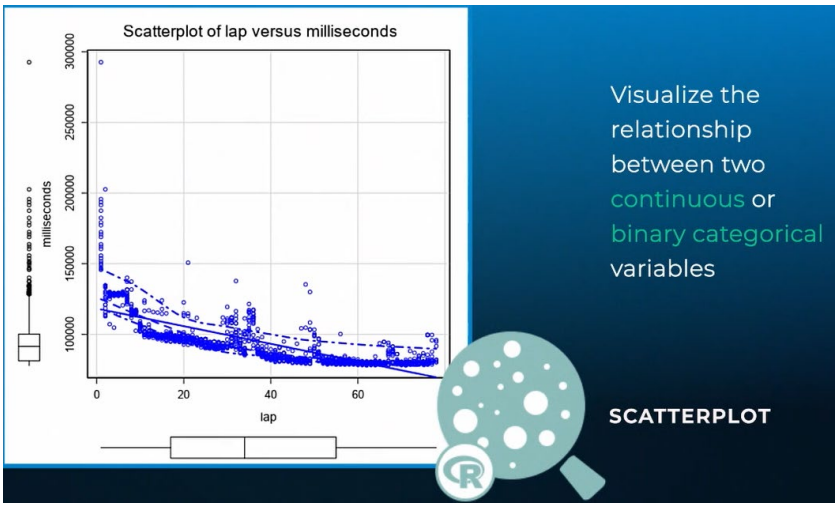
The *Frequency Table* tool is used to get counts of values within a field. However, this tool does not accept the following data types:

a. Fixed Decimal
b. Float
c. Double
d. Date/Time
e. Blob
f. Spatial Object



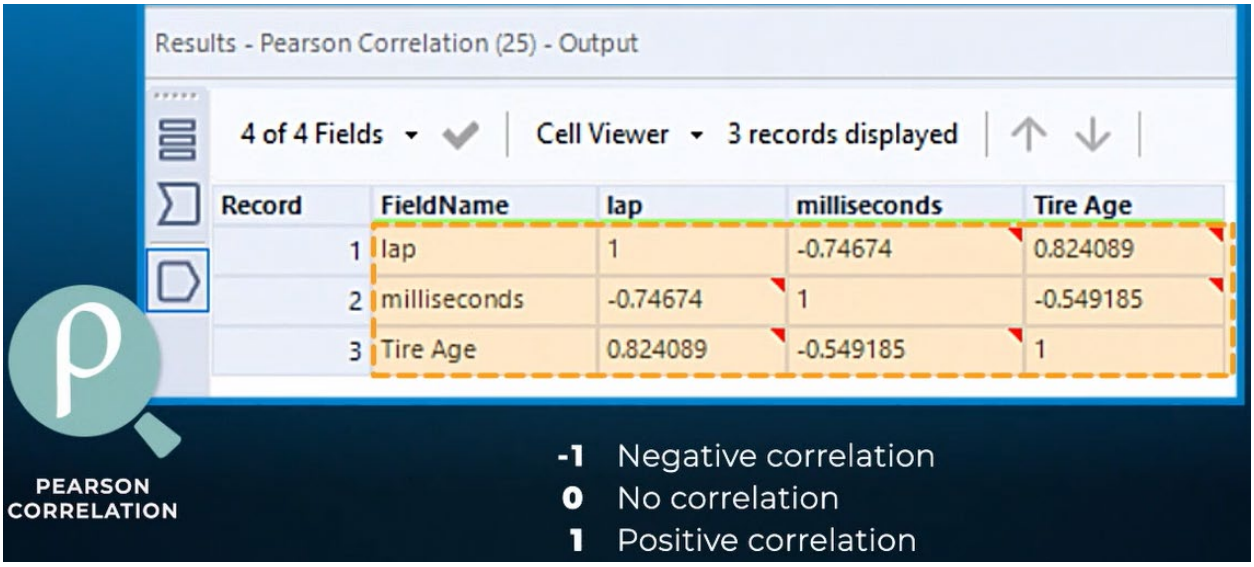
Record	Field_Name	Field_Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	Tire Choice	U	336	24.76	336	24.76
2	Tire Choice	W	336	24.76	672	49.52
3	Tire Choice	I	288	21.22	960	70.74
4	Tire Choice	SS	251	18.50	1211	89.24
5	Tire Choice	S	145	10.69	1356	99.93
6	Tire Choice	[Null]	1	0.07	1357	100.00

The *Scatterplot* tool is useful for visualizing the relationship between two continuous or binary categorical variables. Scatterplots do not necessarily prove a relationship.

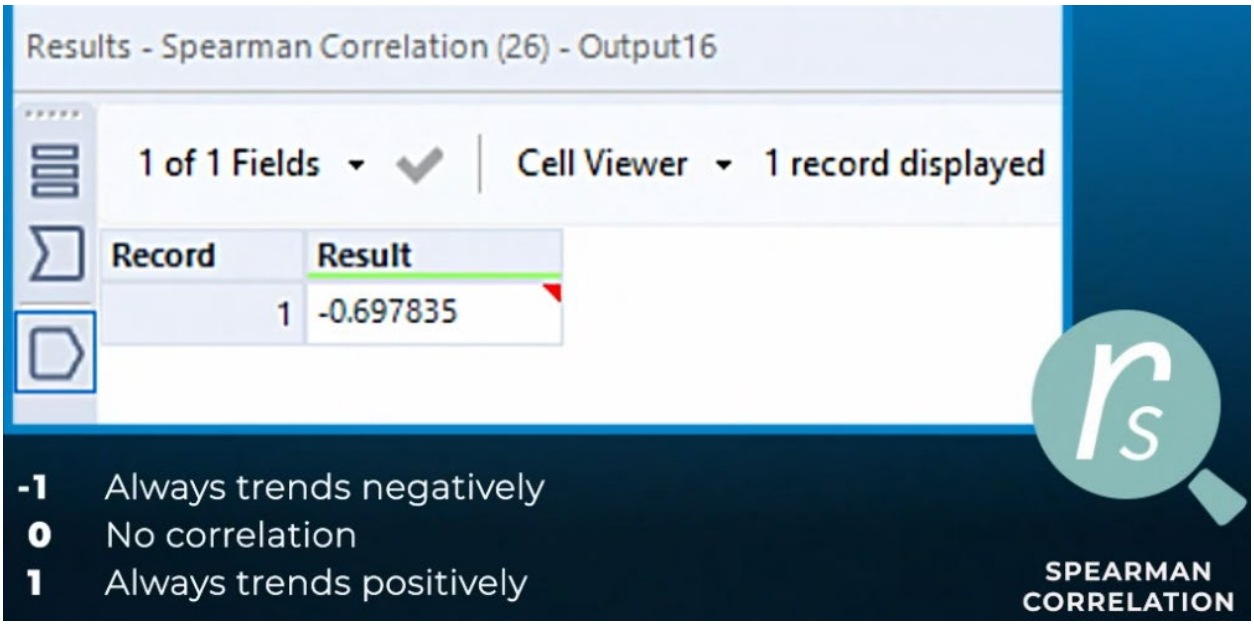


The *Pearson Correlation* and *Spearman Correlation* tools provide a way to see relationships between two continuous variables.

The *Pearson Correlation* tool measures the strength of a linear association between two variables.



The *Spearman Correlation* tool evaluates the monotonic relationship between two variables. The variables must be continuous (i.e., any value within a range, including fractions and decimals) or ordinal (e.g., first, second, third, etc.). If ordinal, the values must be ranked so the order of the relationship can be determined. For example, we can't say that *green* is a step up from *red*. However, *large* is greater than *small*.



The *Association Analysis* tool creates a full correlation matrix of numeric variables.

Question 1

Which of these would be useful for determining the correlation between the two variables in the table extract below. Select all that apply.

- ☐ Pearson Correlation
- ☐ Spearman Correlation
- ☐ Association Analysis
- ☑ None of These

Size	Length
Green	12.2
Purple	15.4
Gold	19.6

Question 2

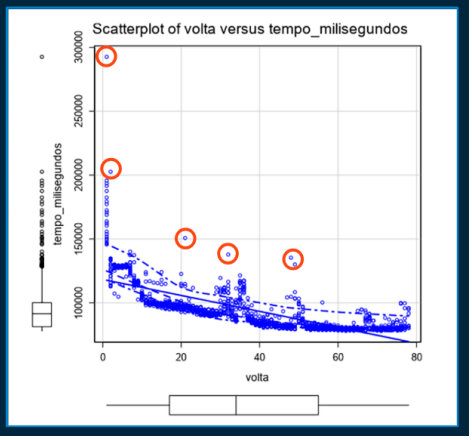
The Field Summary tool works with variables of the following datatypes: *(select all that apply)*

- ☑ String
- ☑ Double
- ☑ Spatial
- ☑ Int16
- ☑ DateTime

Question 3

What do these dots represent on a scatterplot?

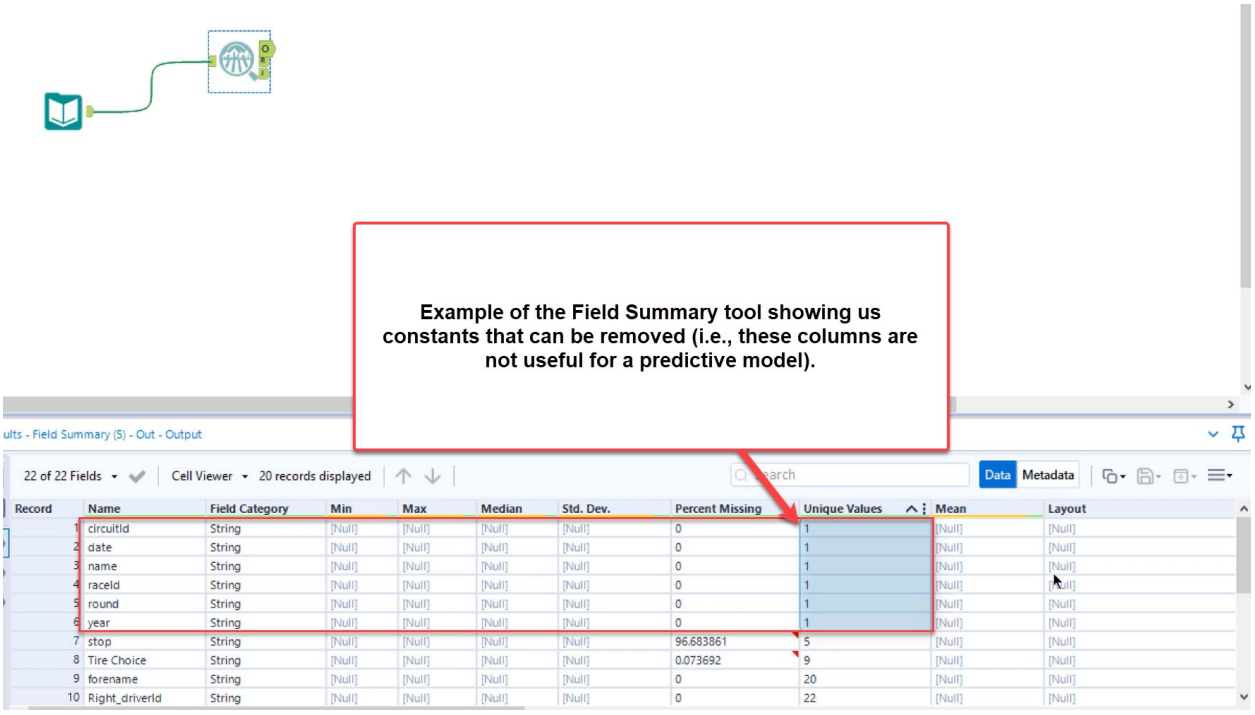
- To be continued
- Not enough datapoints to perform calculations
- Statistically Significant datapoints
- Outliers

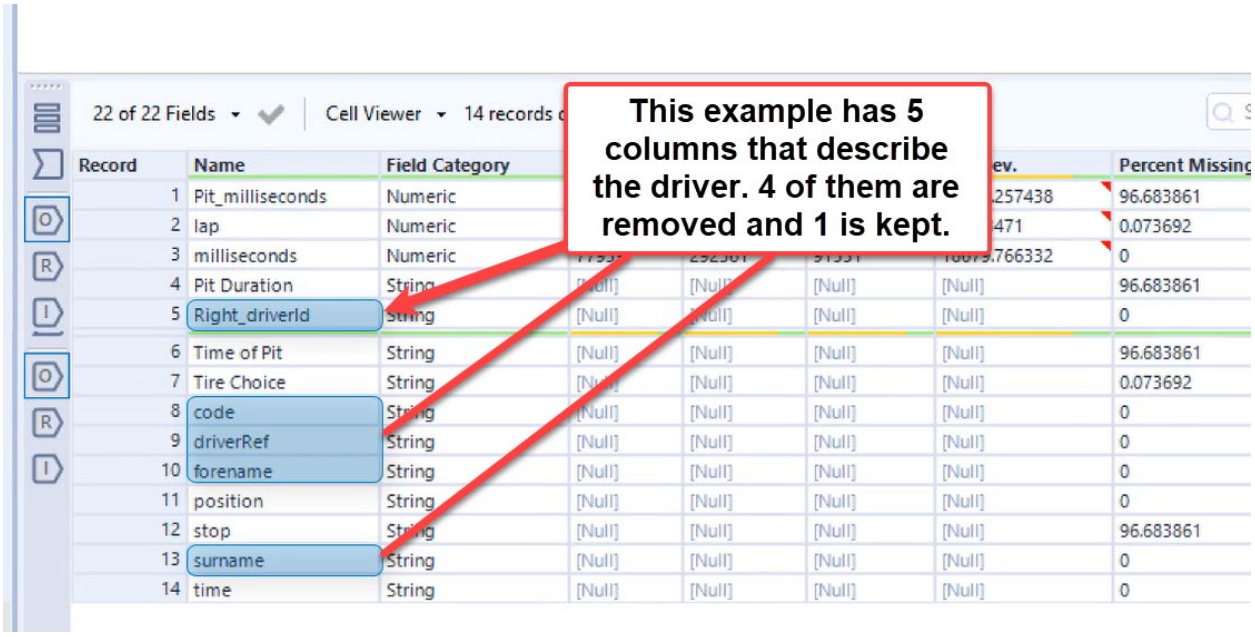


Data Investigation Techniques

Some data provide no predictive value.

Constants
&
Duplicates





Generally speaking, an *outlier* is a value that is more than 2 standard deviations from the mean.

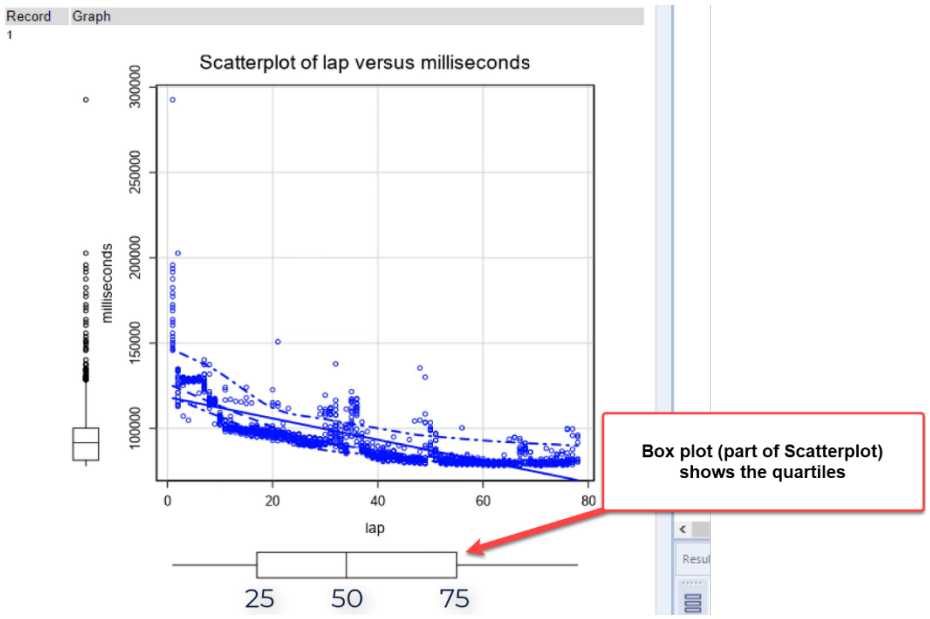
Outliers

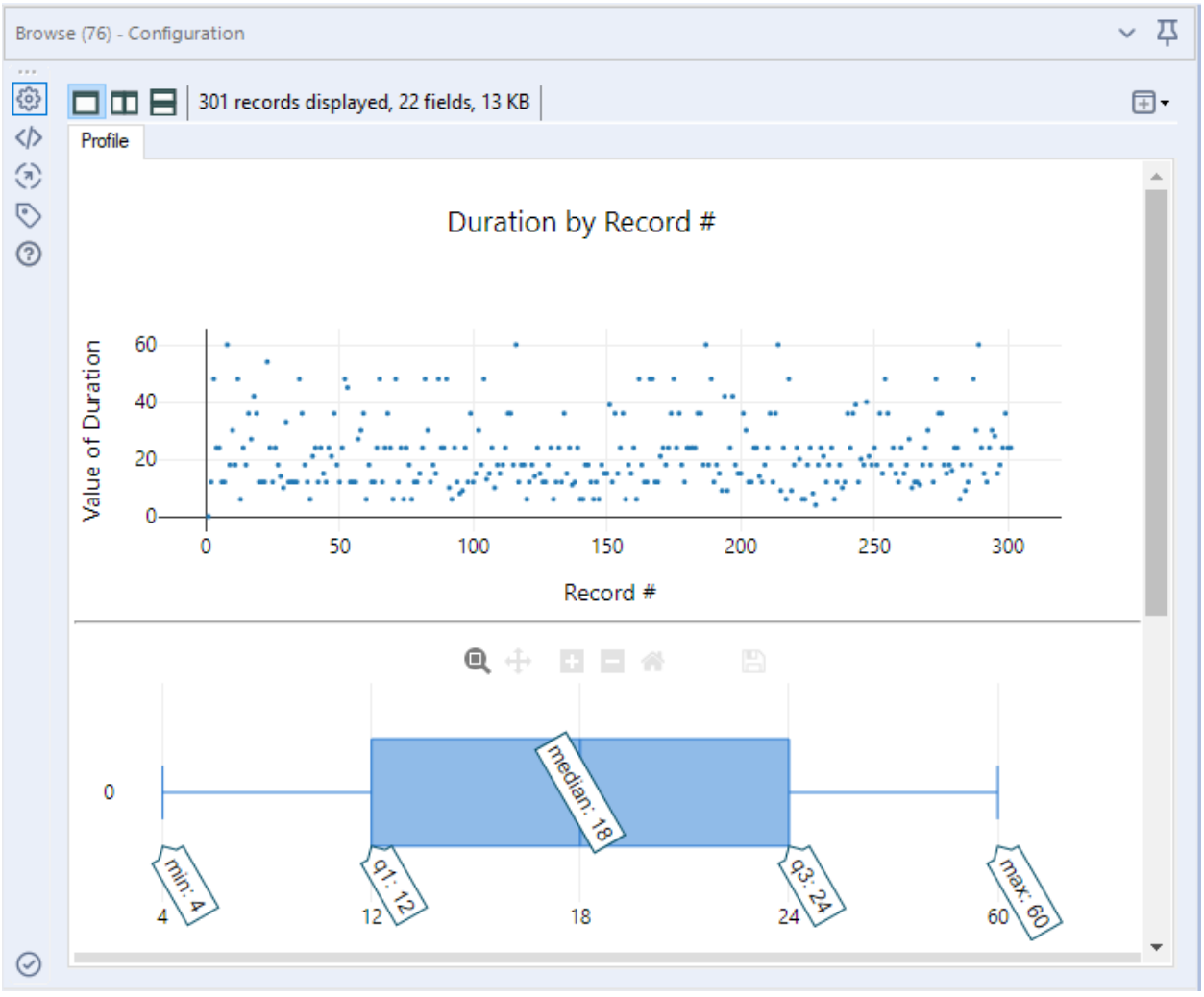
The designation of outlier is subjective and calculations can vary.

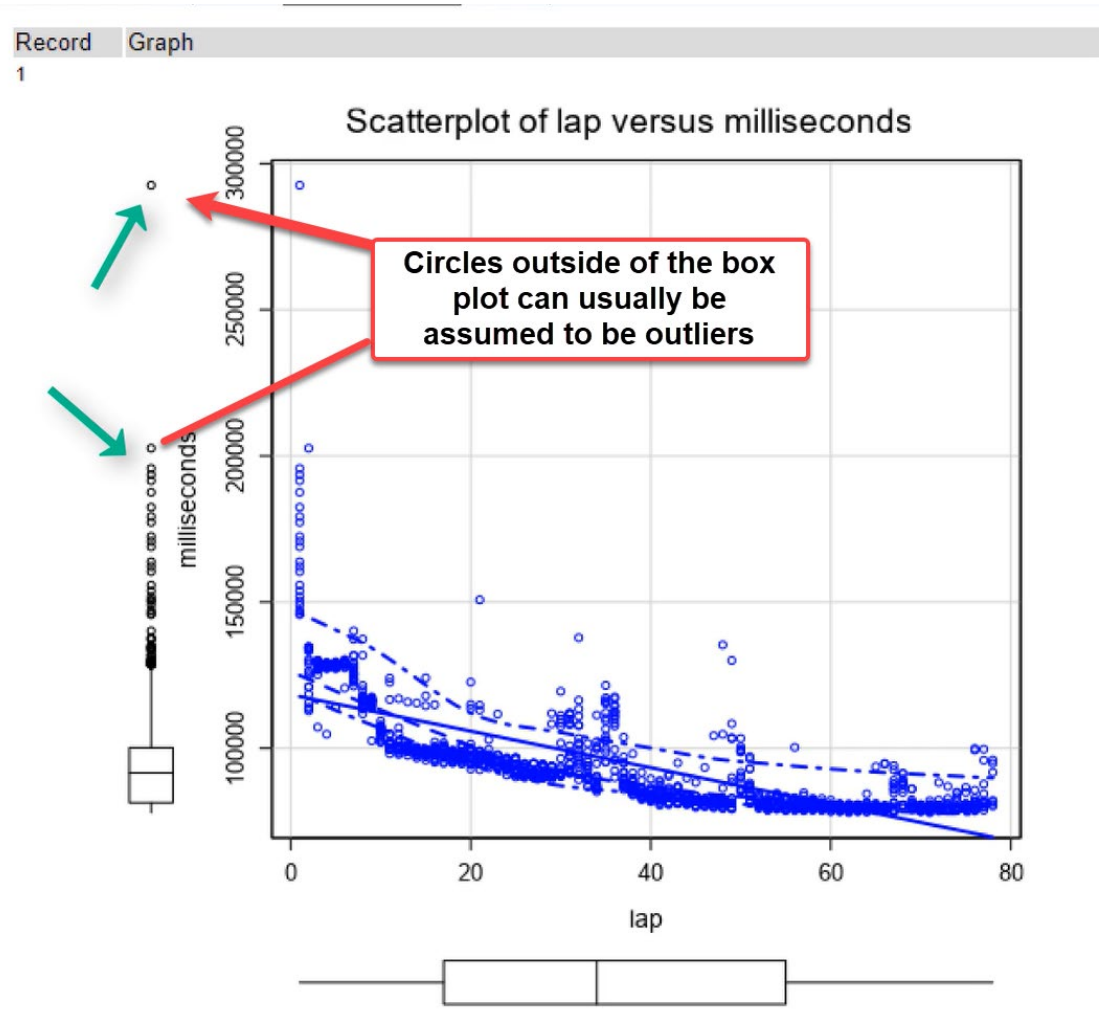
The origin of outliers should be considered before removal.

Outliers can represent niche cases or errors in the dataset.

Do not remove outliers simply because your model cannot explain them.

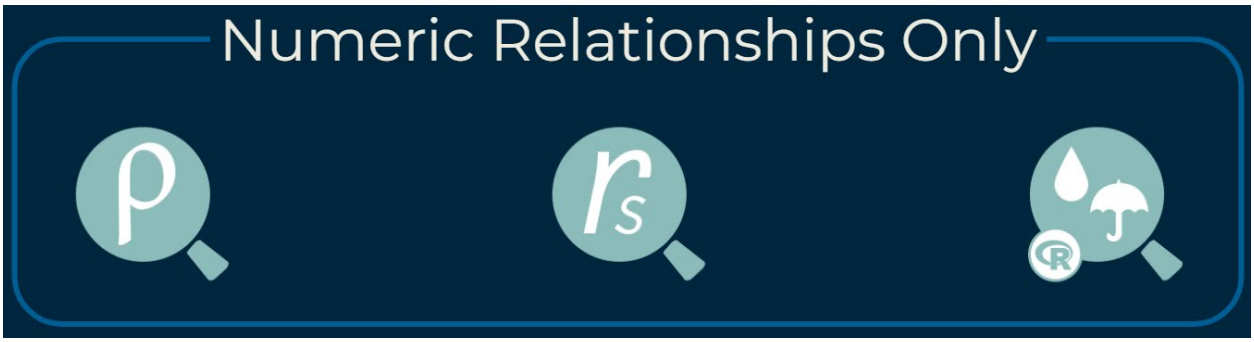

Record
Graph
1
Scatterplot of lap versus milliseconds
milliseconds
300000
250000
200000
150000
100000
0
20
40
60
80
lap
25
50
75
Box plot (part of Scatterplot) shows the quartiles


Browse (76) - Configuration
301 records displayed, 22 fields, 13 KB
Profile
Duration by Record #
Value of Duration
60
40
20
0
0
50
100
150
200
250
300
Record #
0
median: 18
min: 4
q1: 12
q3: 24
max: 60
4
12
18
24
60

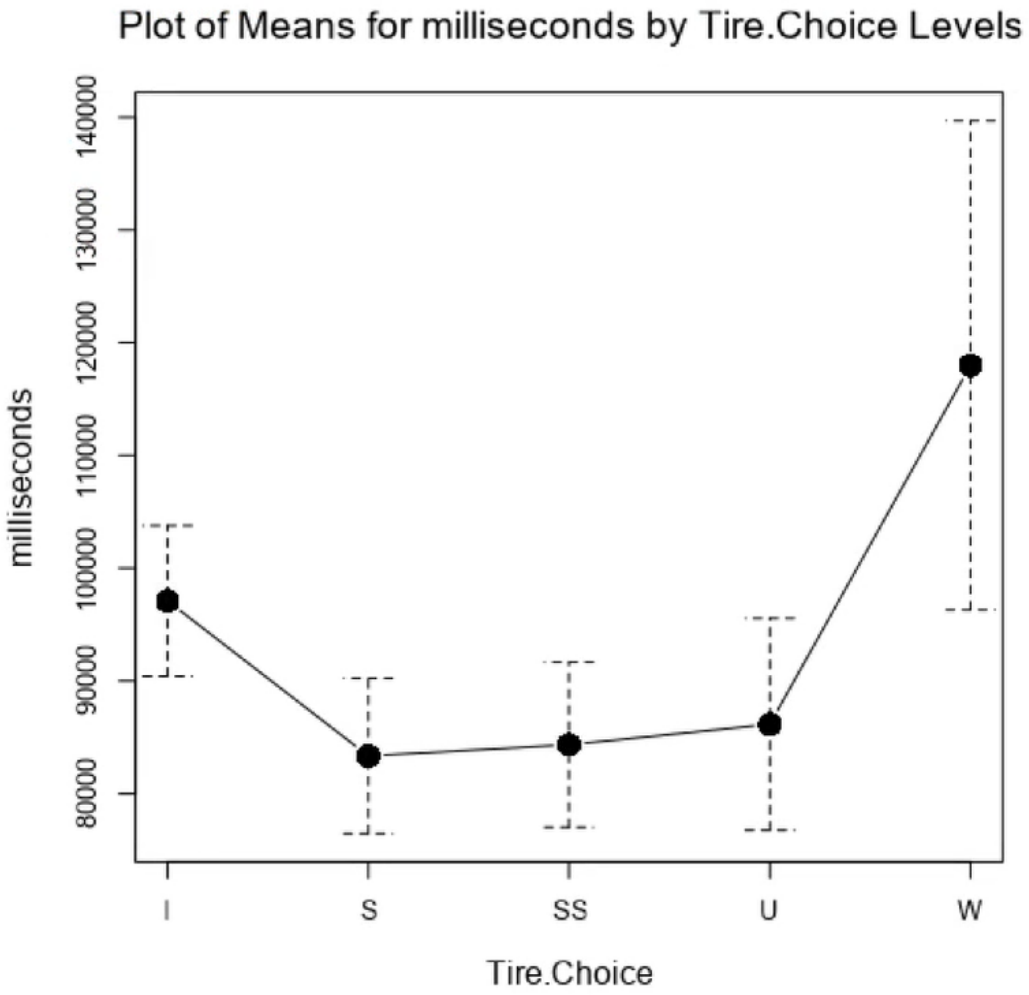


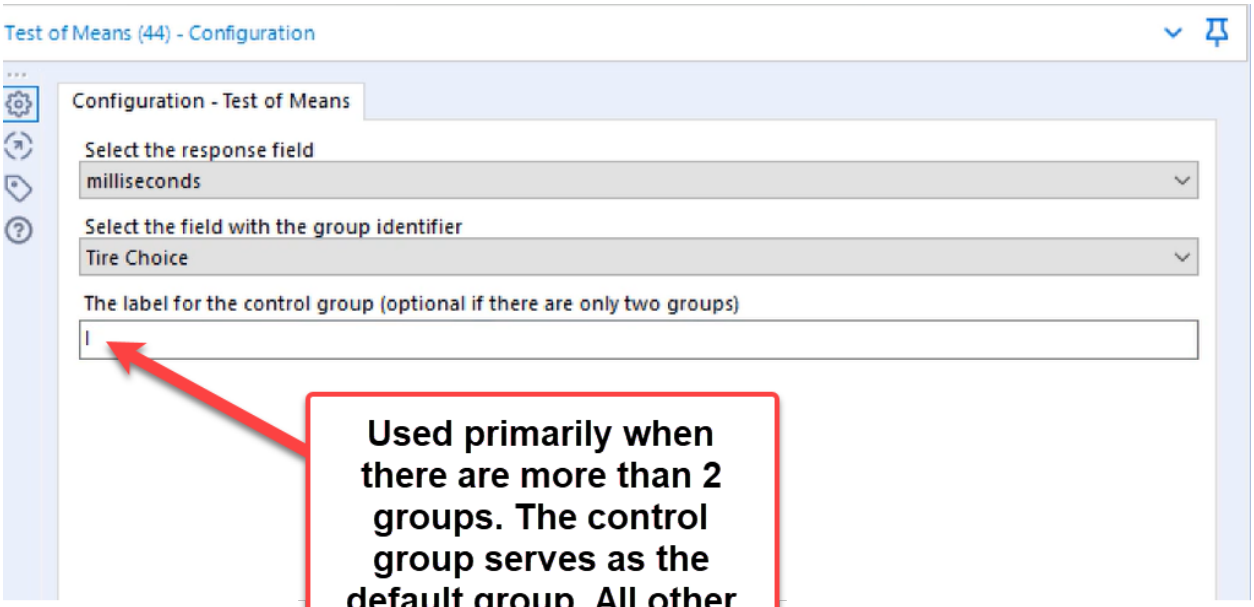
Correlation

- always **two numeric** variables
- can help to identify predictor variables
- can help to reduce collinearity



Plot of Means (a more visual result) and *Test of Means* (a more concrete result) tools are used to evaluate relationships between continuous and categorical variables. An example from the Plot of Means tool is shown below, where the milliseconds for each tire choice is shown, with *standard deviation* selected for the error bars.

Record	Graph
1	

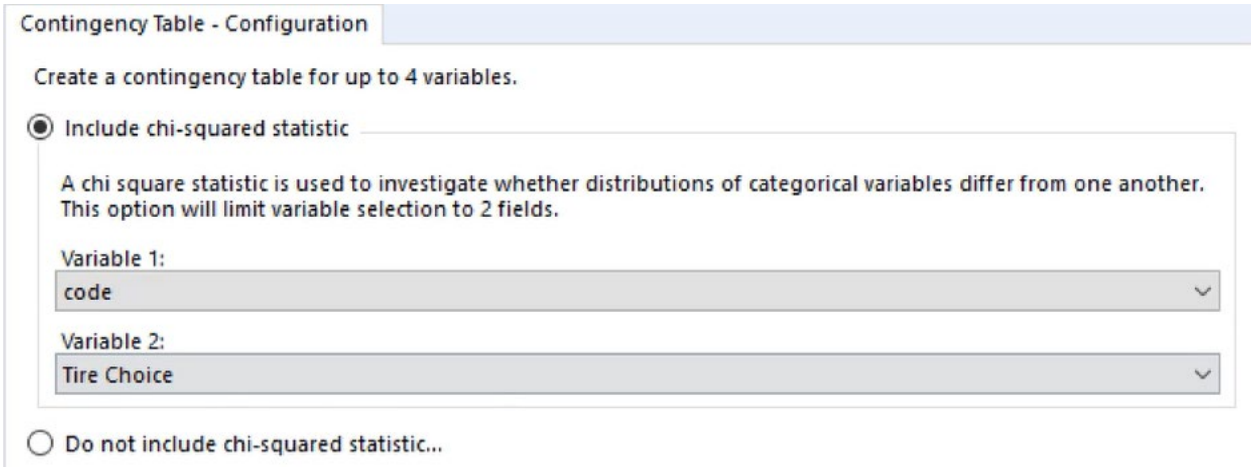


Report | Profile

1 of 1 Fields | Records 1 to 2

Record	Report
1	**Welch's Two Sample t-test(s) of milliseconds by Tire.Choice**

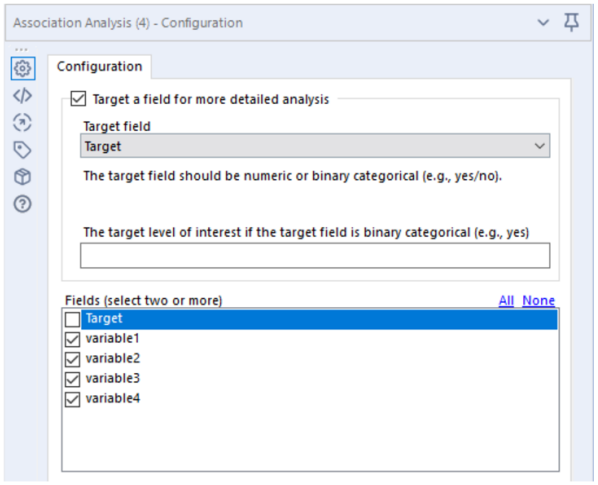
Test	t-Statistic	Degrees of Freedom	p-Value
I vs W	-16.7554	407.45	2.5967e-48
I vs SS	21.015	510.48	4.1459e-71
I vs U	16.9392	603.75	6.0113e-53
I vs S	19.825	281.55	2.3548e-55



Most data scientists spend the majority of their time in the *Data Investigation* phase of the *Data Science Lifecycle* because there's no substitute for knowing your data.

Do multiple fields measure the same thing (multicollinearity)?

Use the Association Analysis Tool to check for this issue and remove redundant fields. Be sure to include the target field you are trying to predict as that will provide more detailed results.



The R output anchor has a Full Correlation Matrix report. A correlation score near 1.0 shows that both predictor columns measure the same thing for predicting the target variable. Similarly, a score near -1.0 shows a high negative correlation (inversely proportional fields).

In this example, variable two and variable four have a correlation score of 1.0, the highest correlation score possible. Use only one of the fields and deselect the other.

Full Correlation Matrix

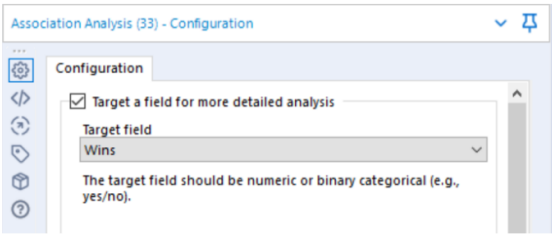
	Target	variable1	variable2	variable3	variable4
Target	1.00000	0.53281	0.48685	0.47953	0.48685
variable1	0.53281	1.00000	0.22644	0.20432	0.22644
variable2	0.48685	0.22644	1.00000	0.25534	1.00000
variable3	0.47953	0.20432	0.25534	1.00000	0.25534
variable4	0.48685	0.22644	1.00000	0.25534	1.00000

The interactive I output anchor shows all fields on both the X-axis and Y-axis. When selecting a box in the grid and the matching X-axis and Y-axis column names will appear along with a pop-up containing the correlation score. Highly positive and negative correlations become a darker color as the correlation increases.

Are the predictor columns statistically significant with a p-value of 0.05 or less?

In simple terms, the p-value is the percentage of chance that any observed correlation between the predictor field and the target field is just random, and no real correlation is occurring. A predictor field is statistically significant when the p-value is 0.05 less, as there is a low chance that there is no actual correlation.

The option for a target field should be selected and set to match the column that will be predicted.



The column list is in order of significance. Check the stars to the right of the p-value column for an easy way to determine if the columns are statistically significant. Fields with a p-value of 0.001 or less receive three stars, 0.01 or less receives two stars, and 0.05 or less has a single star.

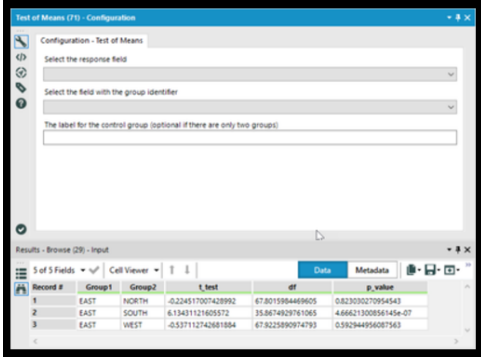
Focused Analysis on Field Wins

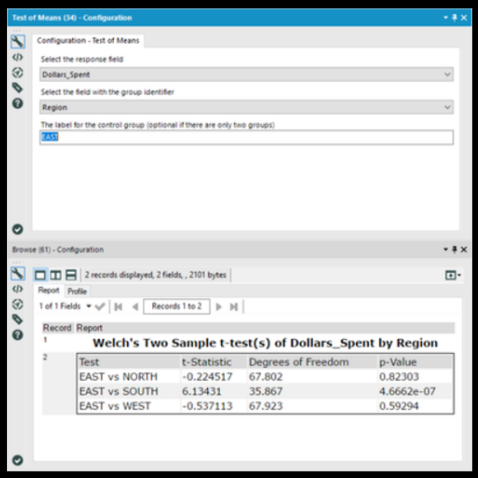
	Association Measure	p-value	
R	0.611273	0.00033261	***
R_G	0.611184	0.00033346	***
OPS	0.593718	0.00054285	***
RBI	0.592518	0.00056077	***
SLG	0.590389	0.00059383	***
TB	0.564118	0.00116689	**
HR	0.513510	0.00370453	**
OBP	0.483960	0.00673336	**
OPS_Adj	0.480936	0.00713741	**
BA	0.467195	0.00924198	**
H	0.438750	0.01528981	*
PA	0.404420	0.02664748	*
BatAge	0.377717	0.03960284	*
IBB	0.335156	0.07021967	.
BB	0.294317	0.11439644	

A common task that analysts can run into (and a good practice when analyzing data) is to determine if the means of 2 sampled groups are significantly different. When this inquest arises, the Test of Means tool is right for you! To demonstrate how to configure this tool and how to interpret the results, a workflow has been attached. The attached workflow (v11.7) compares the amount of money that customers spent across different regions in the US. The **Dollars_Spent** field identifies the amount of money an individual spent and the **Region** field identifies the region that the individual resides in (NORTH, SOUTH, EAST, WEST).

First, let's look at the interface for this tool, which is relatively simple:

1. The "Select response field" allows you to choose the field containing the metric you want to compare. (**Dollars_Spent** in the attached workflow)
2. The "Select the field with the group identifier" input allows you to choose the field that identifies the groups of interest. (**Region** in the attached workflow)
3. The "The label for the control group" input allows you to identify the group you would like to use as the control group. (**East** in the attached workflow)





Choosing the control group is of interest when investigating multiple groups. Since the control group in this example is EAST, the results will compare the means of EAST vs. NORTH, EAST vs. SOUTH, and EAST vs. WEST. The following shows the results from this test:

D Output:

R Output:

From this output, we can determine whether the means of two groups represent a statistically significant difference, and which group has a higher mean.

The p-values in the output pictured above reveal that EAST and SOUTH have the only significantly different means, and the sign associated with the t-statistic (t_test & t-statistic) reveals that the mean of EAST is smaller than NORTH and WEST, but larger than SOUTH.

A note on the t-statistic:

1. A negative sign means the control group has a smaller average value.
2. A positive sign indicates that the control group has a larger average value.

`One Hot Encoding` is a pre-processing step that is applied to categorical data, to convert it into a non-ordinal numerical representation for use in machine learning algorithms.

Let's say we have 3 data instances with attributes of `Preferred Programming Language` and `OS of Choice`.

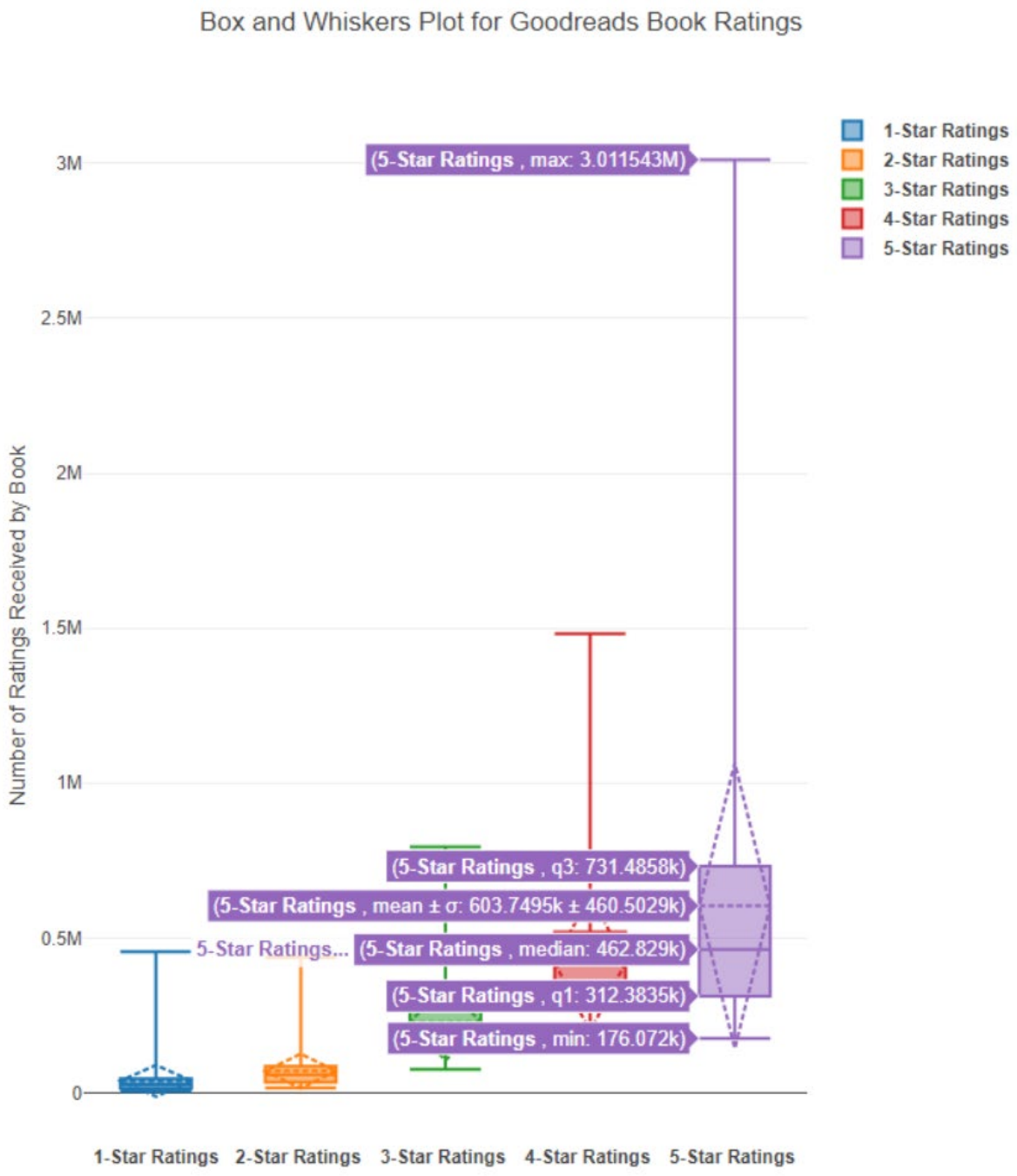
Preferred Programming Language	OS of Choice
Javascript	OSX
Python	Linux
Scala	OSX

A One Hot Encoded version of the Label Encoded table above would look something like this:

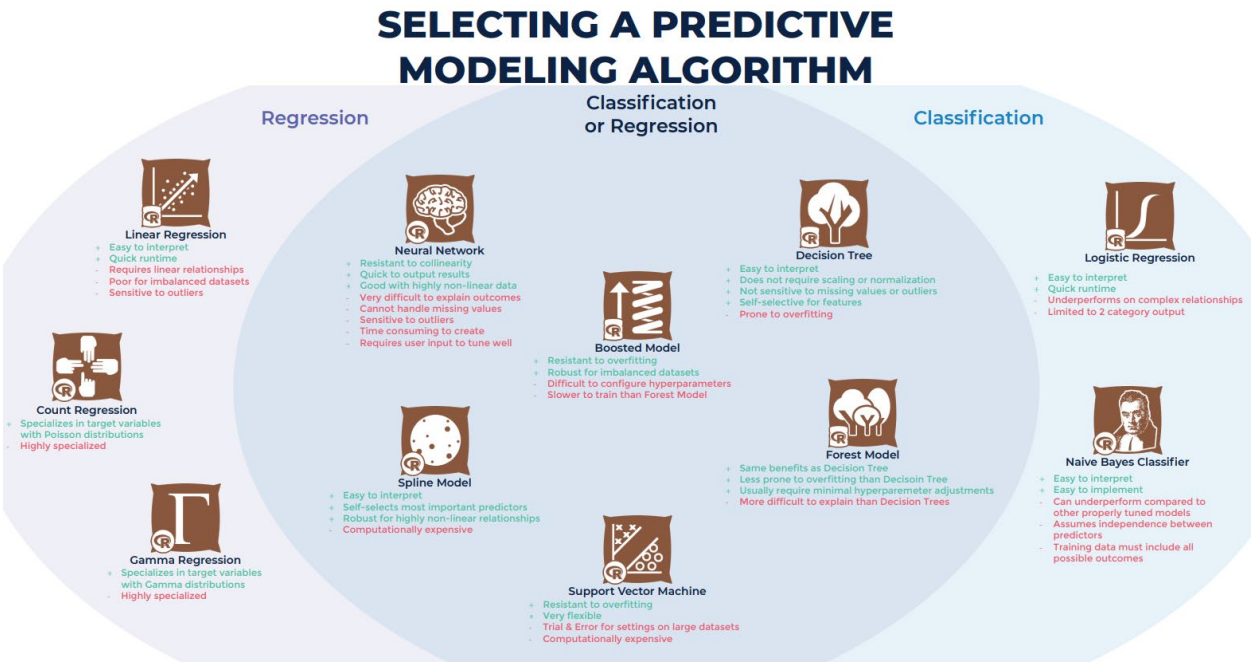
Javascript	Python	Scala	OSX	Linux
1	0	0	1	0
0	1	0	0	1
0	0	1	1	0

If the outlier in question is:

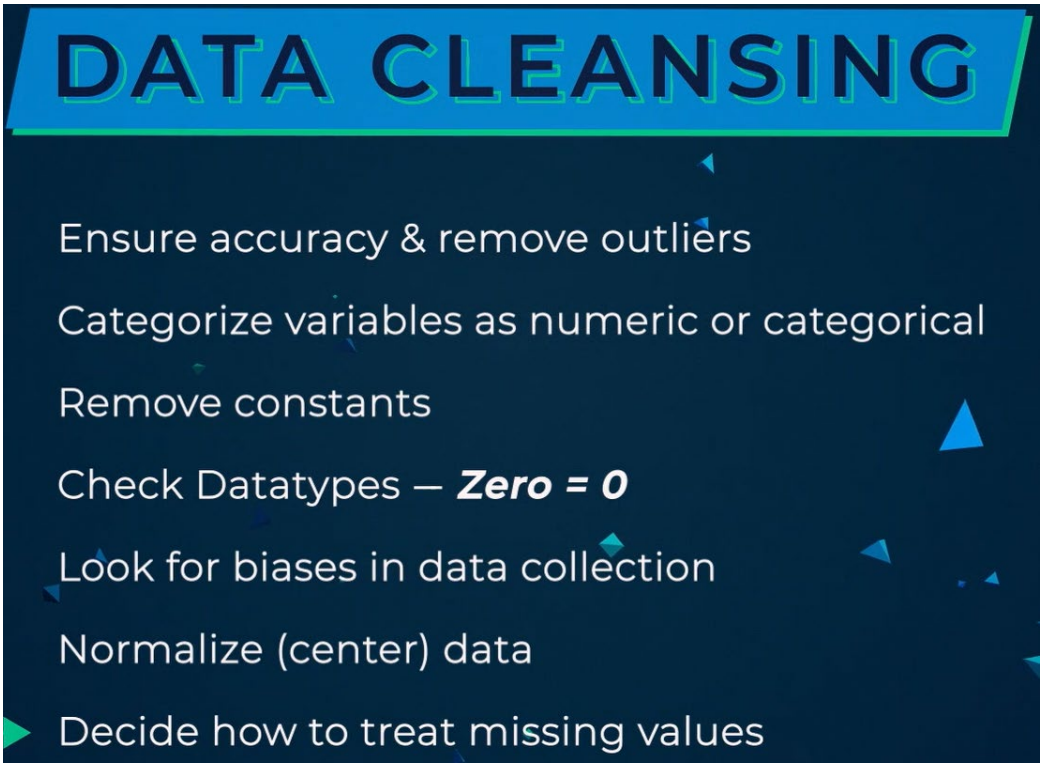
- A measurement error or data entry error, correct the error if possible. If you can't fix it, remove that observation because you know it's incorrect.
- Not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately remove the outlier.
- A natural part of the population you are studying, you should not remove it.

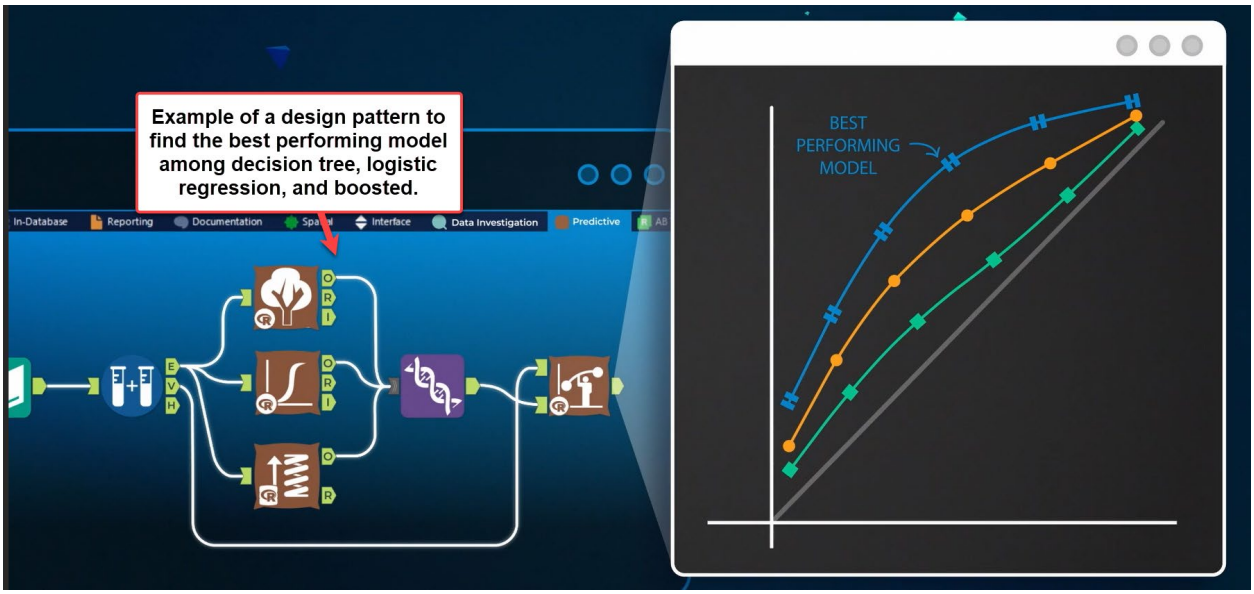


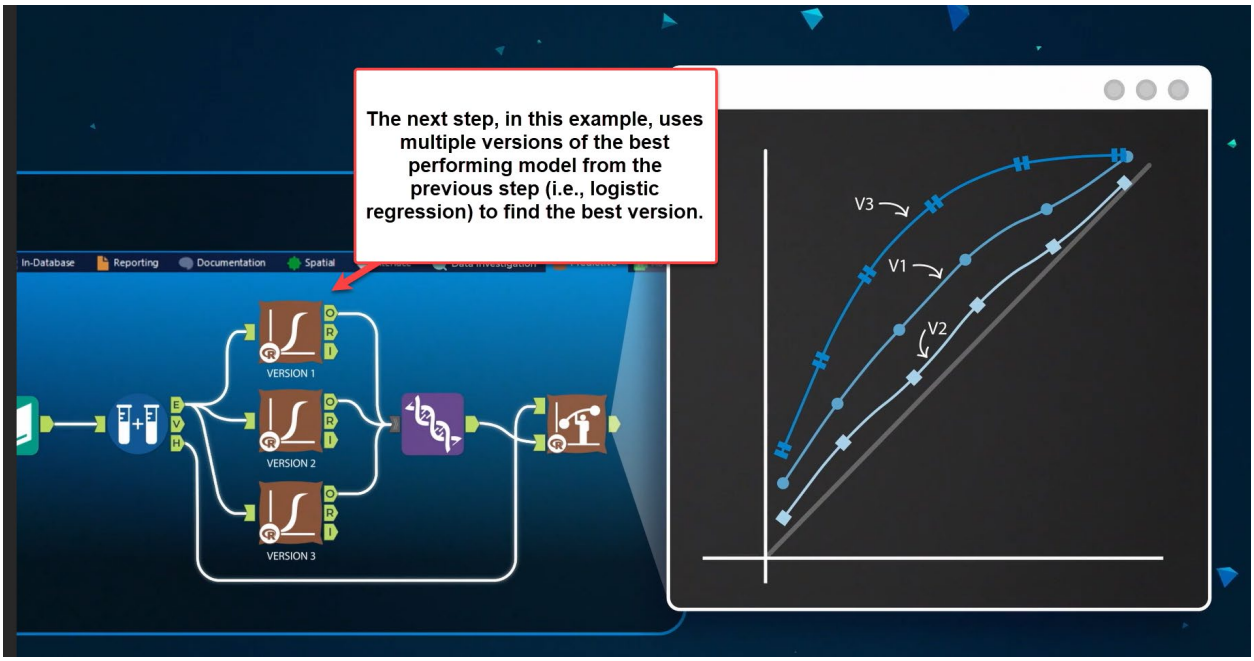
Predictive Modeling

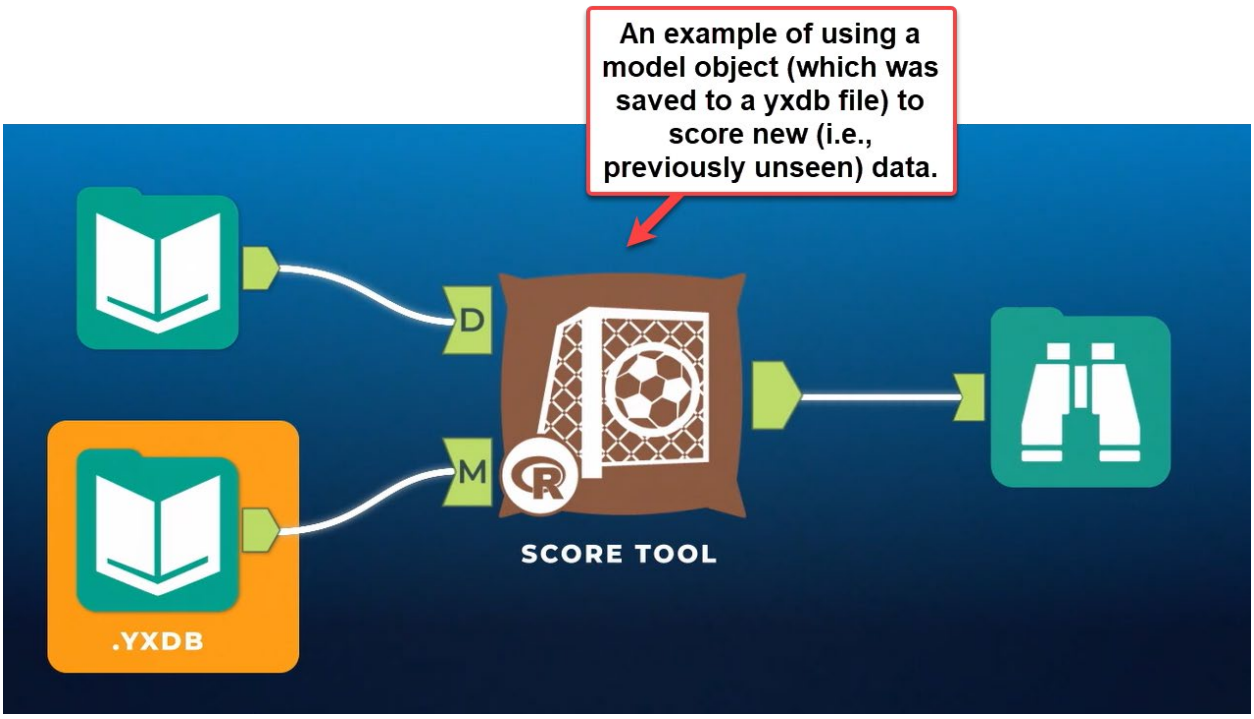


A model is a mapping of the relationships in your data.

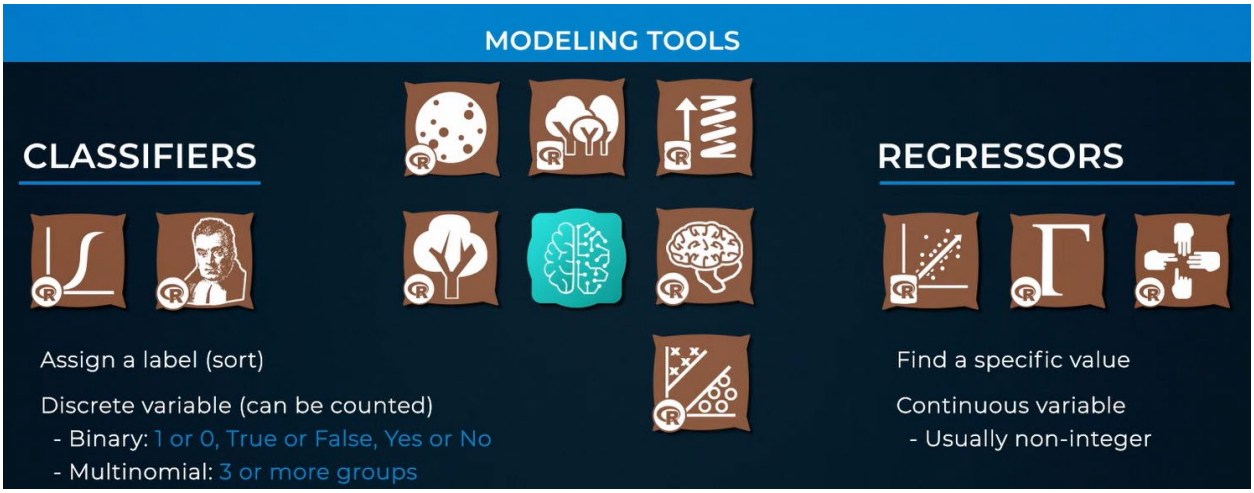


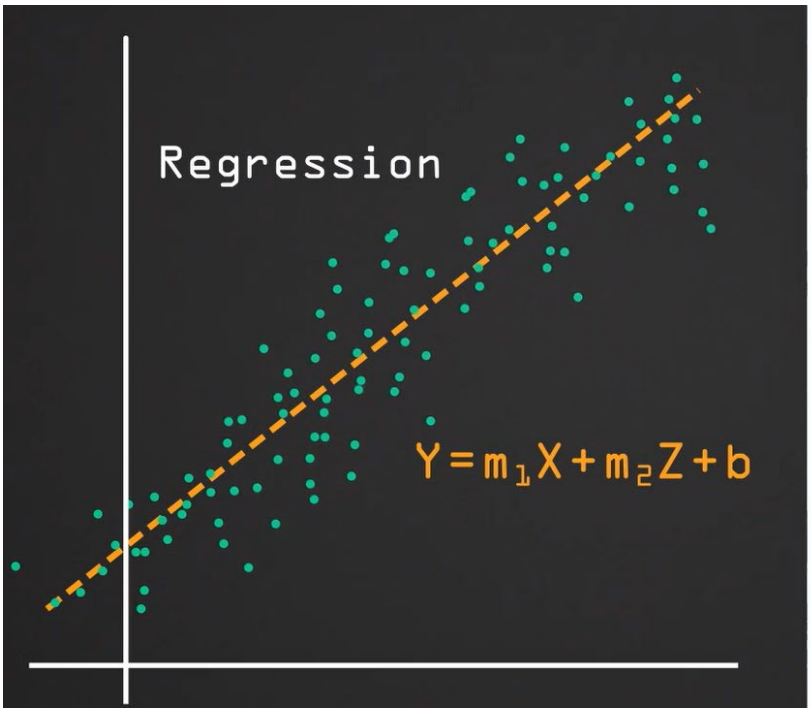

Example of a design pattern to find the best performing model among decision tree, logistic regression, and boosted.
In-Database
Reporting
Documentation
Interface
Data Investigation
Predictive
BEST PERFORMING MODEL


The next step, in this example, uses multiple versions of the best performing model from the previous step (i.e., logistic regression) to find the best version.
In-Database
Reporting
Documentation
Spatial
VERSION 1
VERSION 2
VERSION 3
V3
V1
V2



Predictive Analytics Fundamentals





MEAN

$(\overline{X})$

VARIANCE

$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$

STANDARD DEVIATION

$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}$

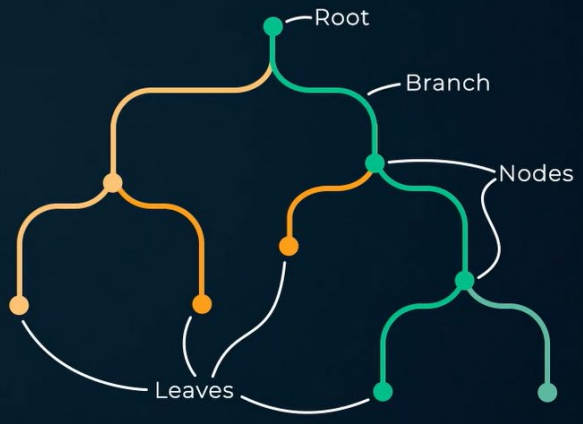
CORRELATION

Relationships between variables

CLASSIFIER

Select the threshold that most accurately separates categories

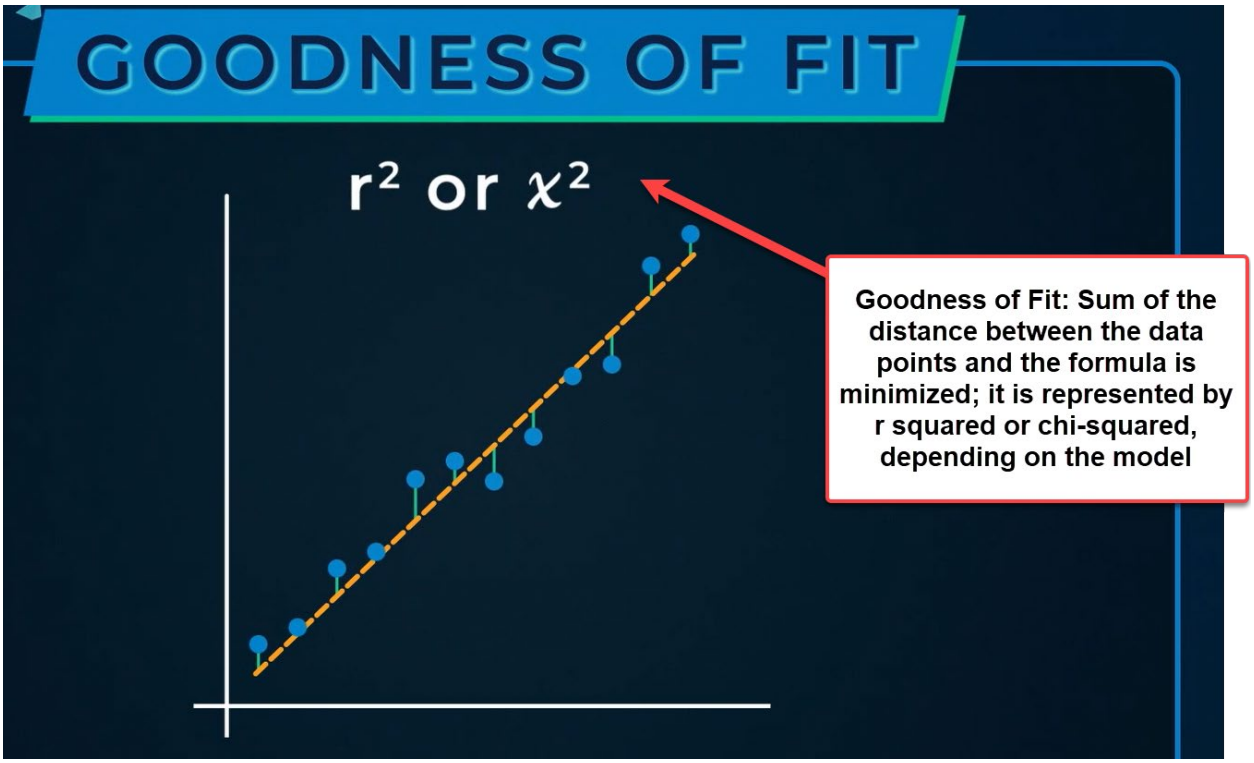
Subject	Category 1	Category 2
A	10	9.0
B	15	8.2
C	8	8.5
D	12	9.8
E	5	8.0
F	18	7.9
G	6	8.3



Question 2

What is collinearity?

- When more than one model fits the data equally well.
- When multiple independent variables describe the same information.
- The relationship between two variables.
- None of these.



An example of splitting data (for model creation) between estimation data and validation data. The estimation data are used to create the model while the validation data are used to assess the suitability of the model

Historical Data

PREDICTOR	PREDICTOR	PREDICTOR	TARGET
1	.7	5	124
3	.6	6	115
5	.5	5	101
3	.5	8	128
4	.4	6	111
5	.6	3	129

Estimation Data

PREDICTOR	PREDICTOR	PREDICTOR	TARGET
1	.7	5	124
3	.6	6	115
5	.5	5	101
3	.5	8	128

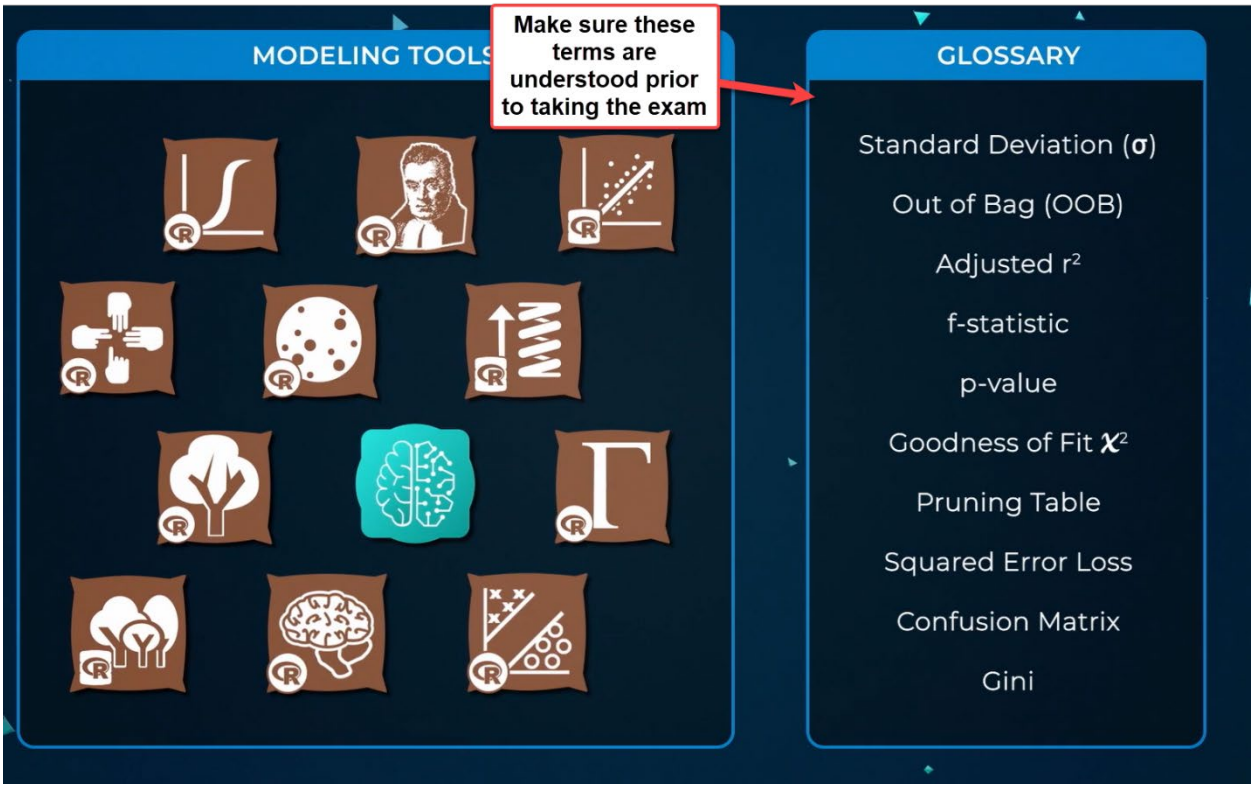
Validation Data

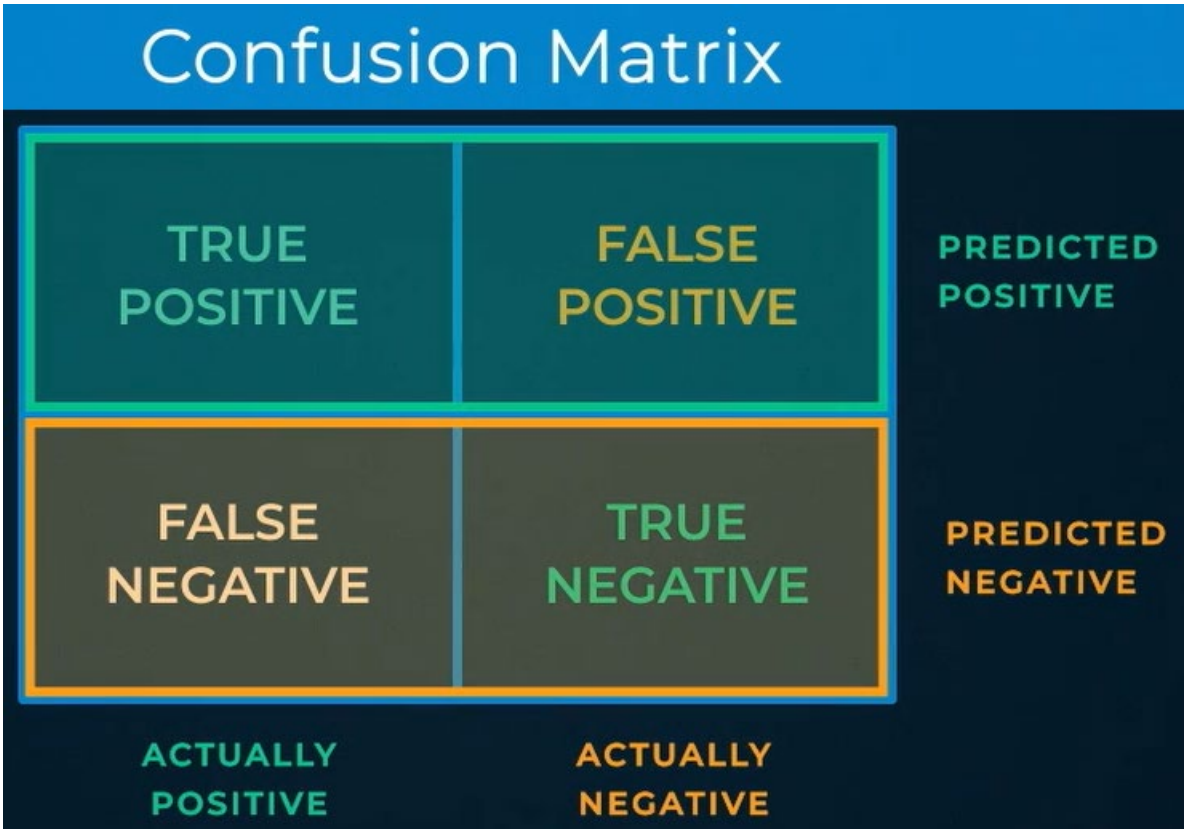
PREDICTOR	PREDICTOR	PREDICTOR	TARGET
4	.4	6	111
5	.6	3	129

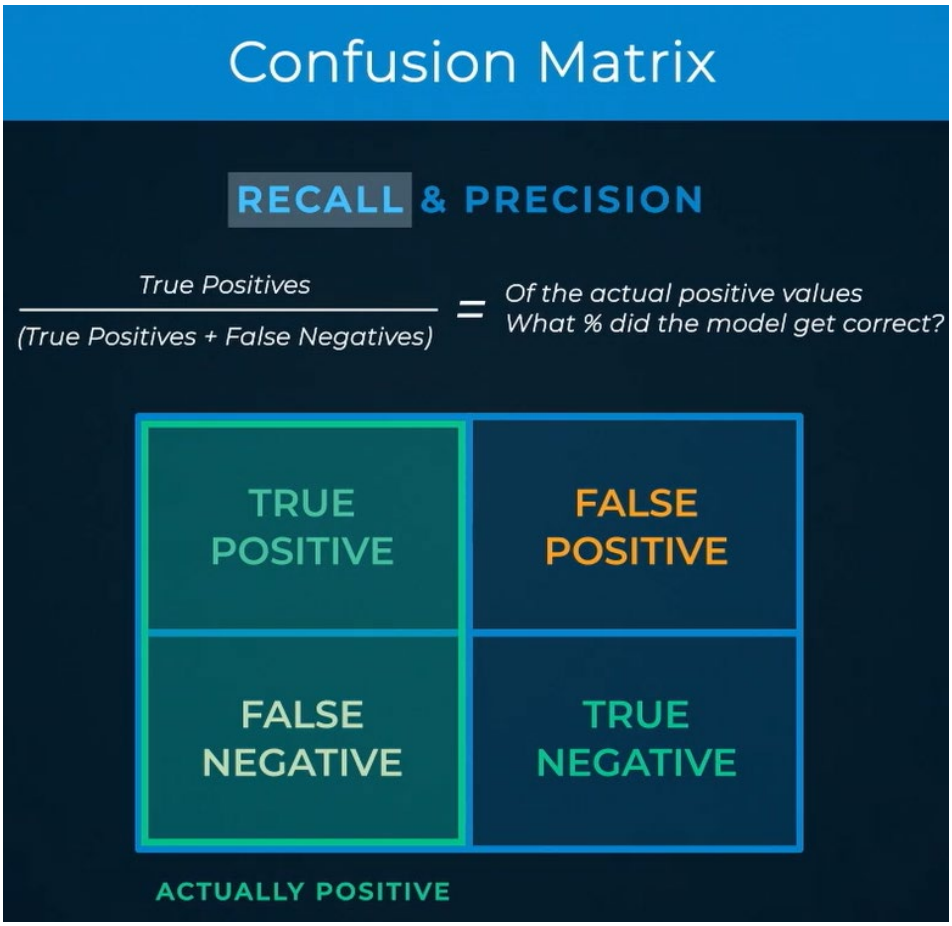
Question 3

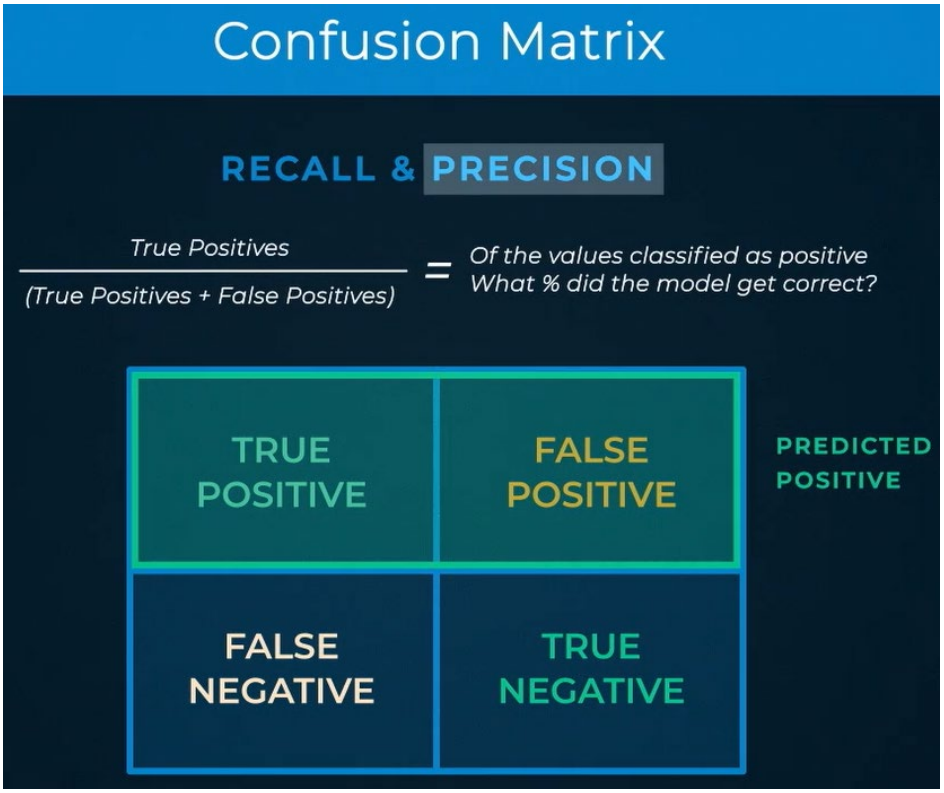
Why is overfitting a problem?

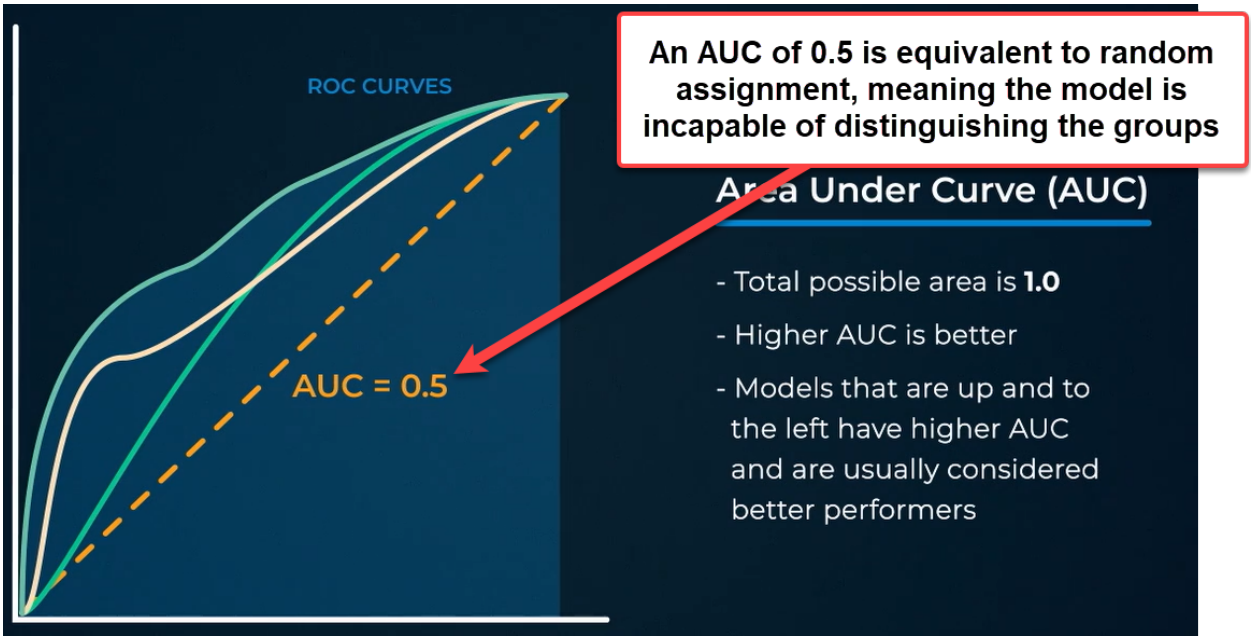
- An overfit model will not generalize to other datasets.
- An overfit model does not appear statistically significant.
- It isn't a problem for all models, just bivariate models.
- Overfitting indicates improper sampling procedure.

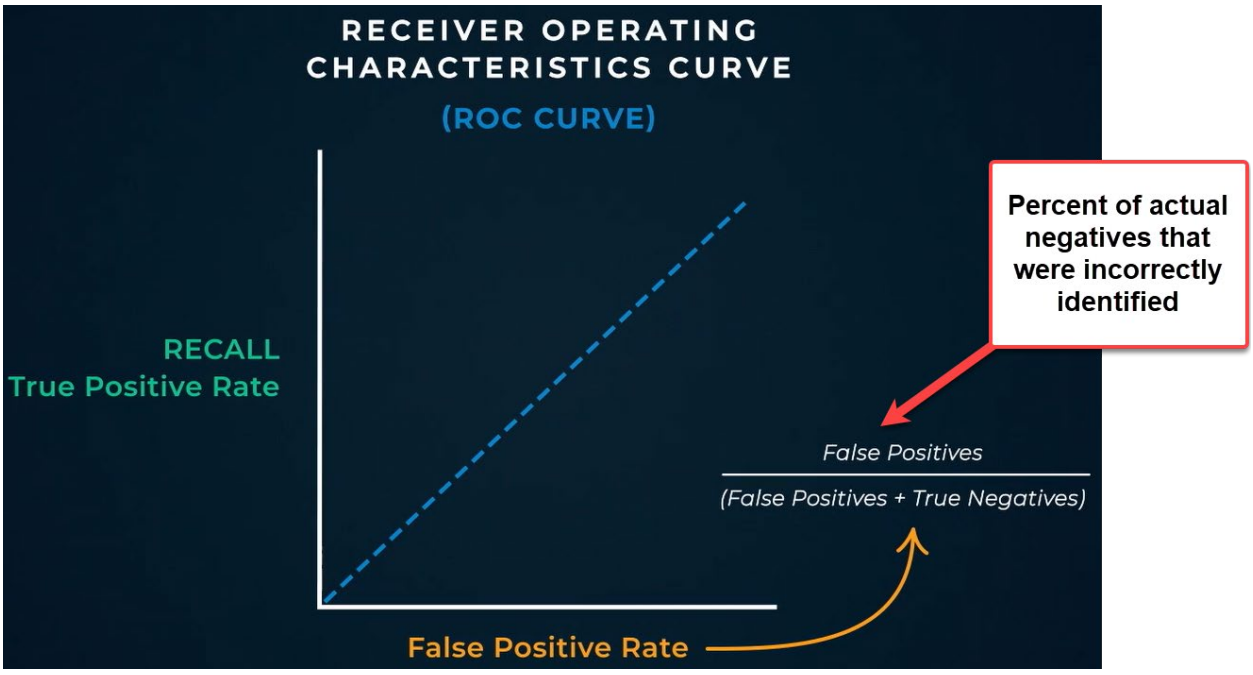



Confusion Matrix
TRUE POSITIVE
FALSE POSITIVE
PREDICTED POSITIVE
FALSE NEGATIVE
TRUE NEGATIVE
PREDICTED NEGATIVE
ACTUALLY POSITIVE
ACTUALLY NEGATIVE


Confusion Matrix
RECALL & PRECISION
True Positives
(True Positives + False Negatives)
=
Of the actual positive values
What % did the model get correct?
TRUE POSITIVE
FALSE POSITIVE
FALSE NEGATIVE
TRUE NEGATIVE
ACTUALLY POSITIVE

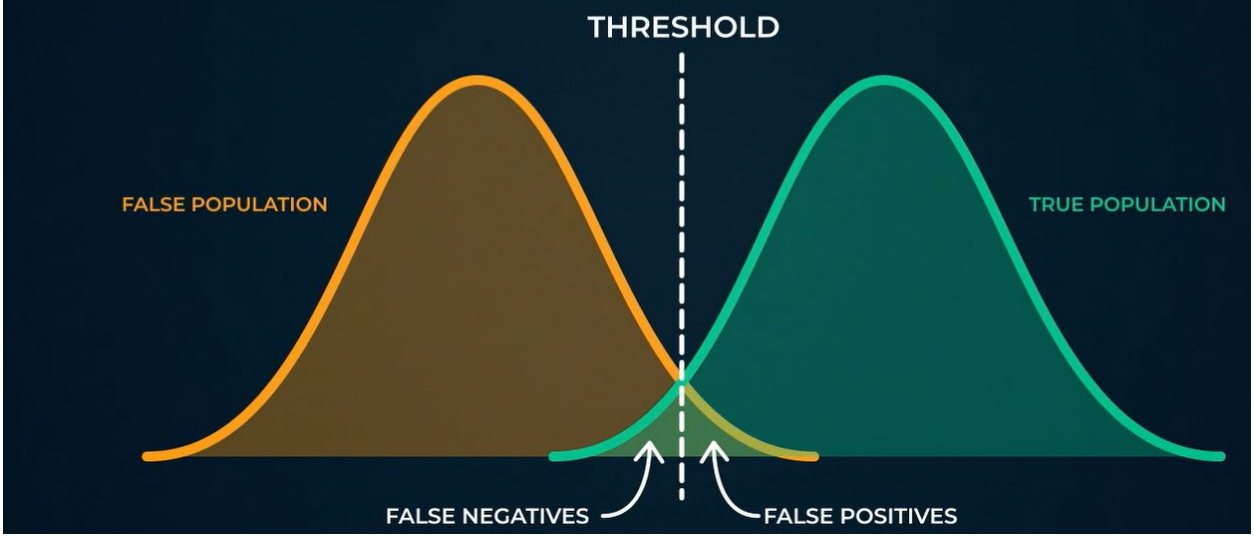

Confusion Matrix
RECALL & PRECISION
True Positives
(True Positives + False Positives)
=
Of the values classified as positive
What % did the model get correct?
TRUE POSITIVE
FALSE POSITIVE
FALSE NEGATIVE
TRUE NEGATIVE
PREDICTED POSITIVE


ROC CURVES
AUC = 0.5
An AUC of 0.5 is equivalent to random assignment, meaning the model is incapable of distinguishing the groups
Area Under Curve (AUC)
- Total possible area is 1.0
- Higher AUC is better
- Models that are up and to the left have higher AUC and are usually considered better performers


RECEIVER OPERATING CHARACTERISTICS CURVE
(ROC CURVE)
RECALL
True Positive Rate
False Positive Rate
False Positives
(False Positives + True Negatives)
Percent of actual negatives that were incorrectly identified

ROC Curve

Visualize model effectiveness across all possible thresholds



Question 4

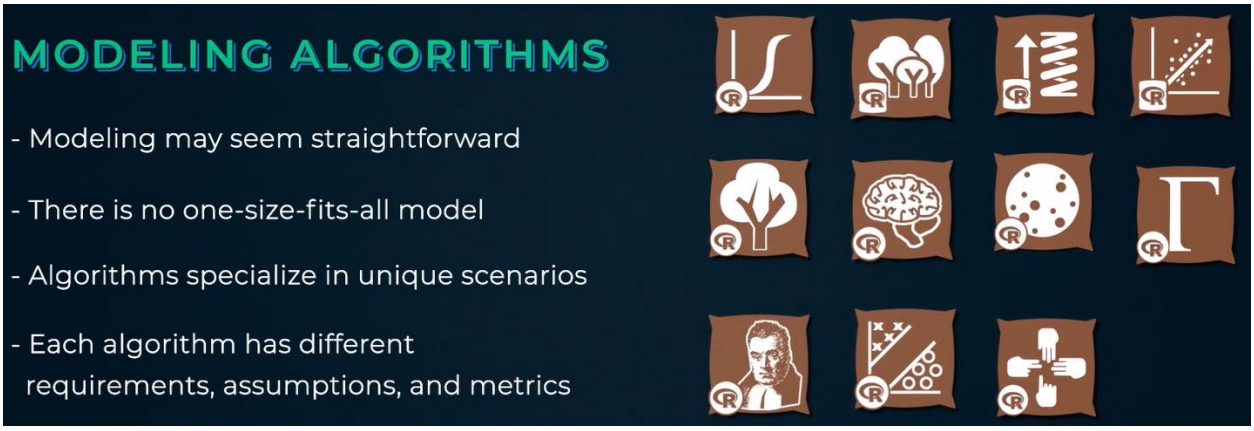
Given the confusion matrix below, how many were predicted positive by the model?

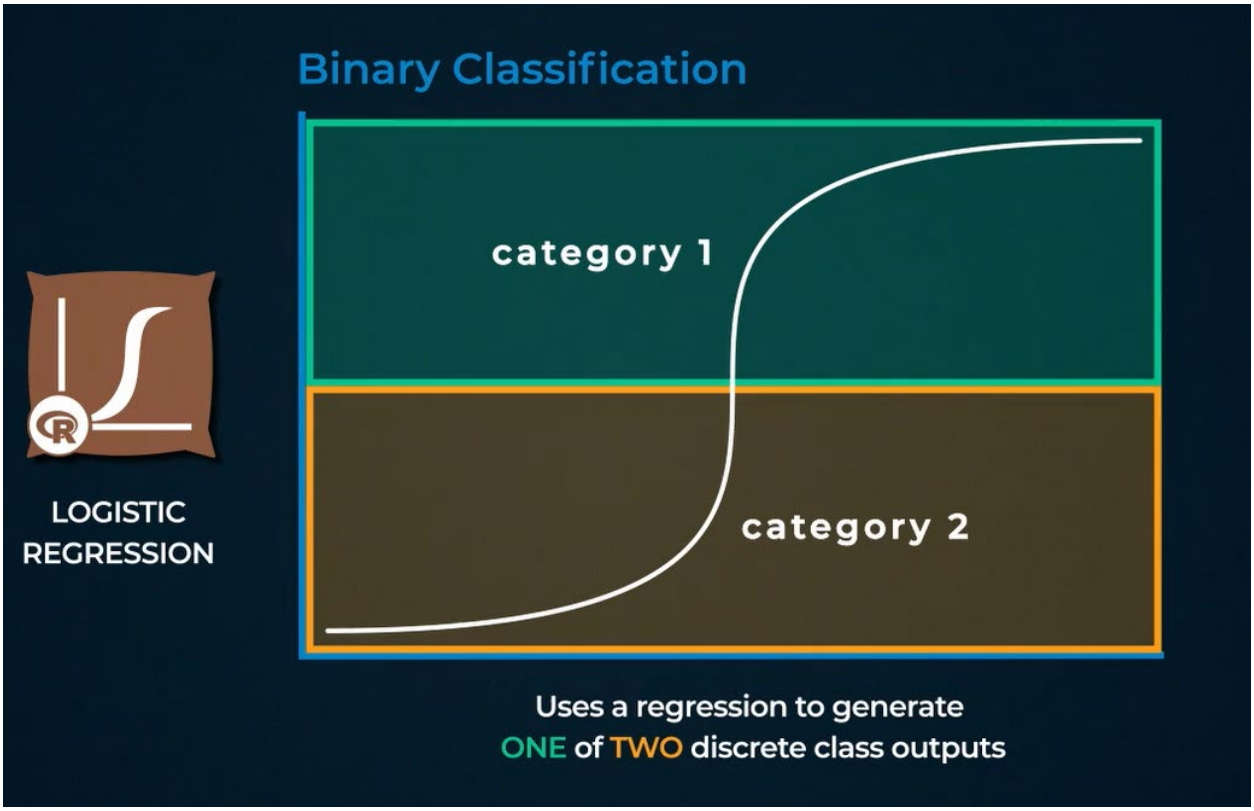
- 55
- 25
- 60
- 51

Confusion Matrix

	Yes	No
Yes	25	30
No	26	35

Creating a Predictive Model

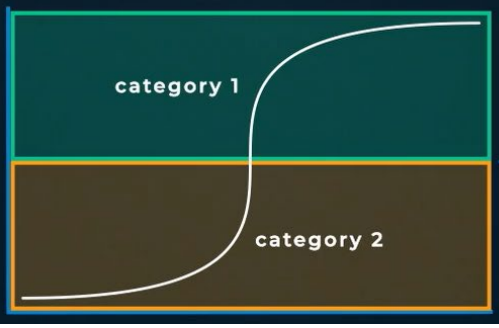




$$\log\left(\frac{p}{1-p}\right) = b_1 x_1 + b_2 x_2 + b_0$$

$$p = \left(\frac{e^{\text{logit}}}{1+e^{\text{logit}}}\right)$$ Probability for each row

LOGISTIC REGRESSION



Threshold for classification
p > threshold = category 1
p < threshold = category 2

LOGISTIC REGRESSION

Report for Logistic Regression Model Logistic

Basic Summary

Call:

glm(formula = Default ~ Num_Loans + Amount + Duration + Age, family = binomial("probit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.959	-1.067	-0.744	1.174	1.694

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.935e-02	0.3705909	-0.2411	0.80948
Num_Loans	-9.450e-02	0.1591620	-0.5937	0.5527
Amount	4.375e-05	0.0000408	1.0722	0.28365
Duration	2.378e-02	0.0098597	2.4122	0.01586 *
Age	-1.405e-02	0.0089571	-1.5688	0.11669

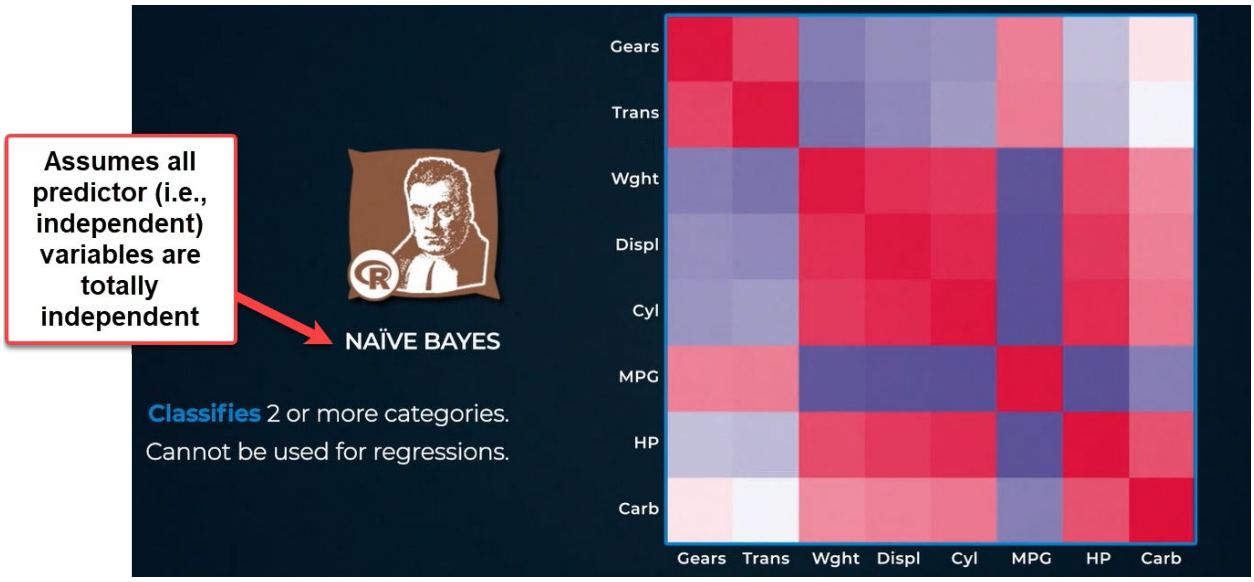
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

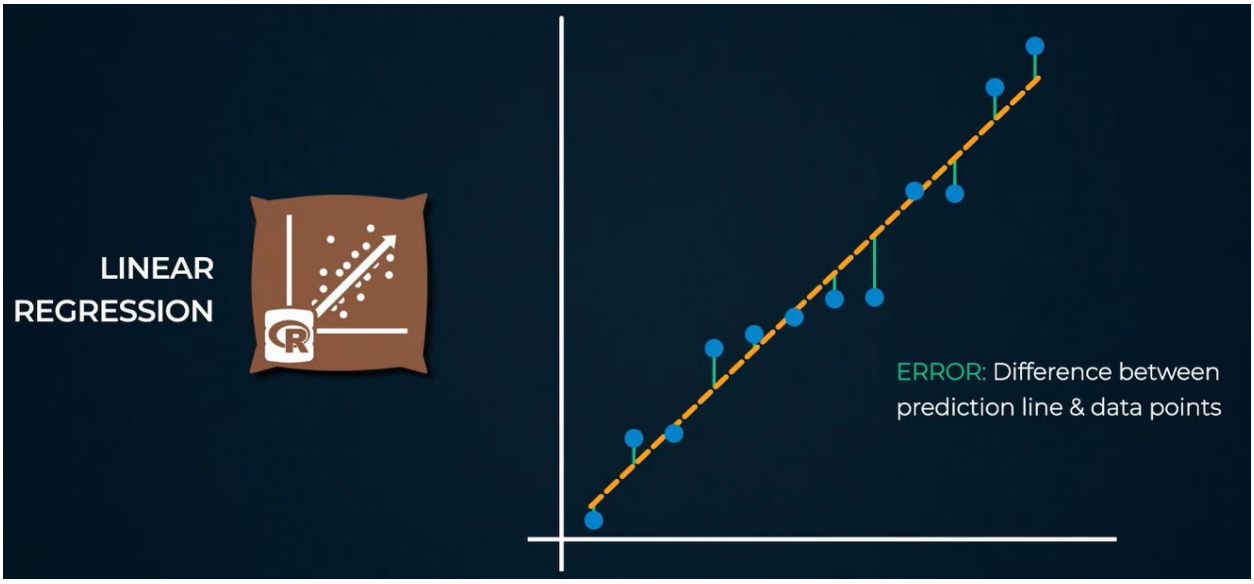
(Dispersion parameter for binomial taken to be 1)

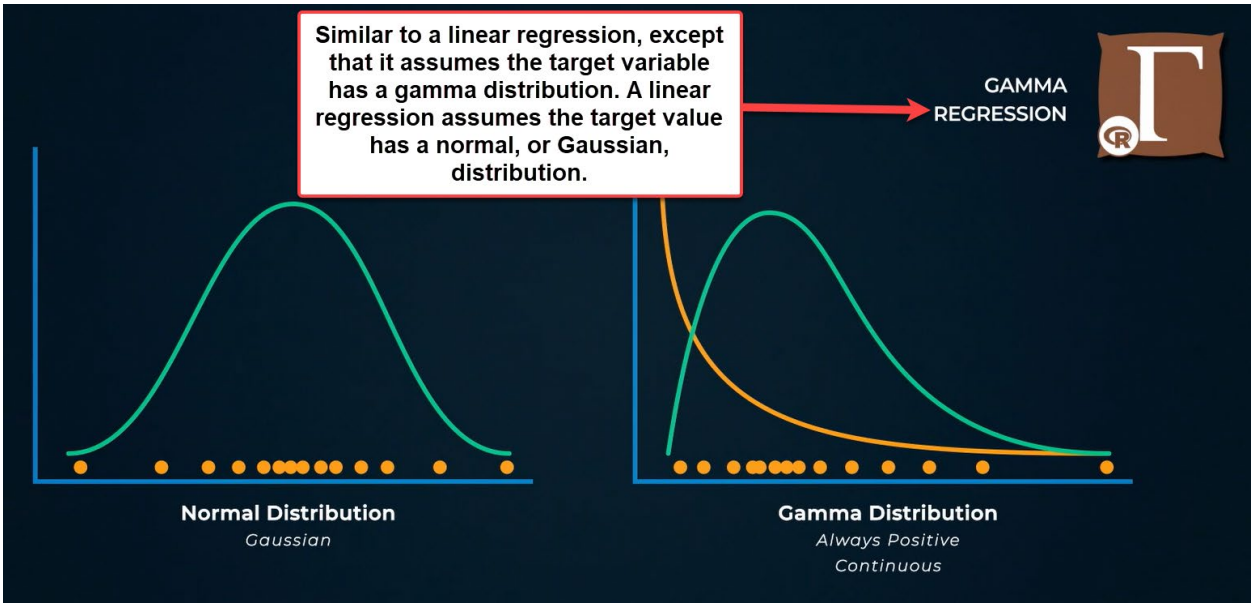
Null deviance: 275.26 on 198 degrees of freedom
Residual deviance: 257.43 on 194 degrees of freedom
McFadden R-Squared: 0.0648, Akaike Information Criterion 267.4
Number of Fisher Scoring iterations: 4
e II Analysis of Deviance Tests

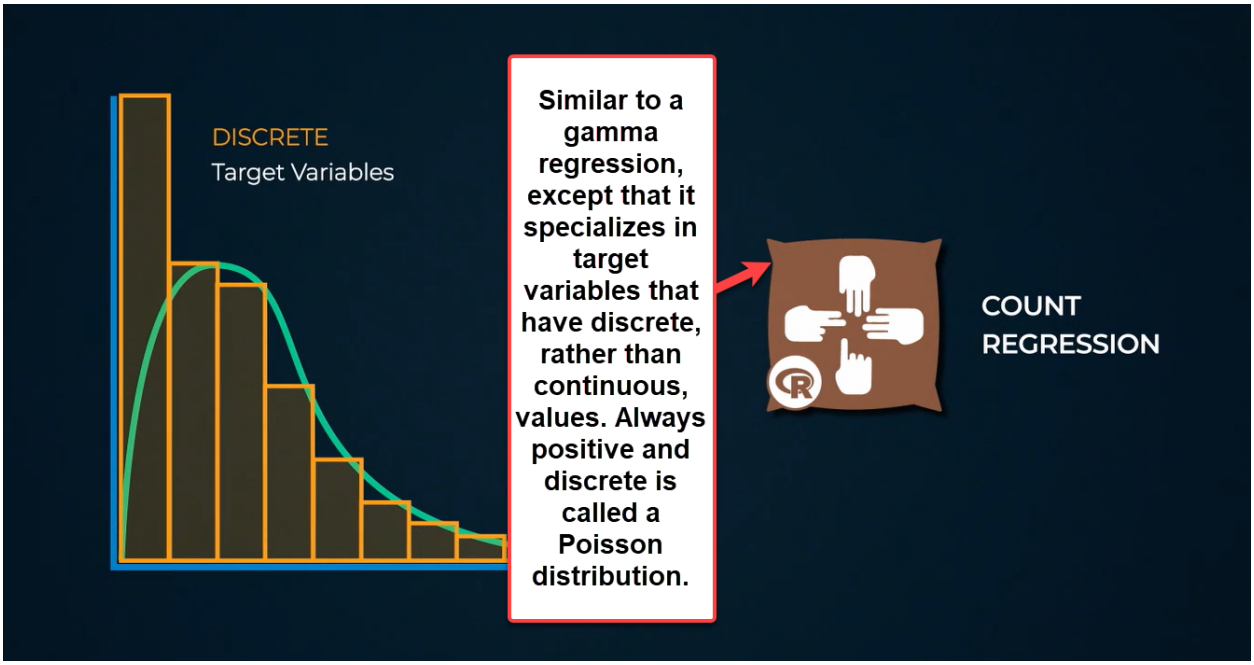
Called a "pseudo" r-squared; serves the same purpose as a traditional r-squared, but can provide lower values

Lower value for Akaike Information Criterion (AIC) is better; this may be used when comparing similar models


Assumes all predictor (i.e., independent) variables are totally independent
NAÏVE BAYES
Classifies 2 or more categories.
Cannot be used for regressions.
Gears
Trans
Wght
Displ
Cyl
MPG
HP
Carb
Gears Trans Wght Displ Cyl MPG HP Carb


LINEAR REGRESSION
ERROR: Difference between prediction line & data points


Similar to a linear regression, except that it assumes the target variable has a gamma distribution. A linear regression assumes the target value has a normal, or Gaussian, distribution.
GAMMA REGRESSION
Normal Distribution
Gaussian
Gamma Distribution
Always Positive
Continuous


DISCRETE
Target Variables
Similar to a gamma regression, except that it specializes in target variables that have discrete, rather than continuous, values. Always positive and discrete is called a Poisson distribution.
COUNT REGRESSION

Summary Report for Decision Tree Model Decision_Tree

Call:

rpart(formula = BuyAComputer ~ AnnualIncome + Age + YearsOfEducation, data = the.data, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 0, usesurrogate = 0, surrogatestyle = 0)

Model Summary

Variables actually used in tree construction:

[1] YearsOfEducation

Root node error: 18/29 = 0.62069

n= 29

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.27778	0	1.00000	1.1111	0.13841
2	0.00000	1	0.72222	1.1111	0.13841

Leaf Summary

node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 29 18 0 (0.37931034 0.34482759 0.27586207)

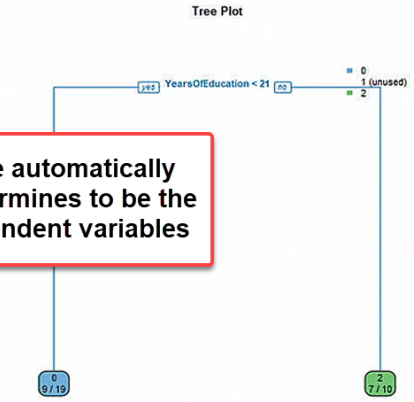
2) YearsOfEducation< 20.5 19 10 0 (0.47368421 0.47368421 0.05263158) *

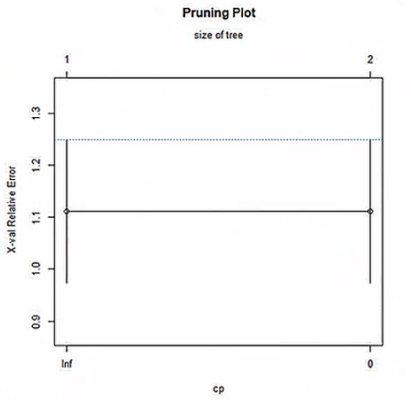
3) YearsOfEducation>=20.5 10 3 2 (0.20000000 0.10000000 0.70000000) *

The Decision Tree automatically selects what it determines to be the appropriate independent variables

Each split is "associated" with a label. Were you to apply those labels after the first grouping, your model would have this accuracy.

Only on Classification (i.e., NOT regression) trees





Summary Report for Decision Tree Model Decision_Tree

Call:

```
rpart(formula = BuyAComputer ~ AnnualIncome + Age +
YearsOfEducation,    data = the.data, minsplit = 20, minbucket = 7, xval =
10,    maxdepth = 20, cp = 0, usesurrogate = 0, surrogatestyle = 0)
```

Model Summary

Variables actually used in tree construction:
[1] YearsOfEducation
Root node error: 18/29 = 0.62069
n= 29

Pruning Table

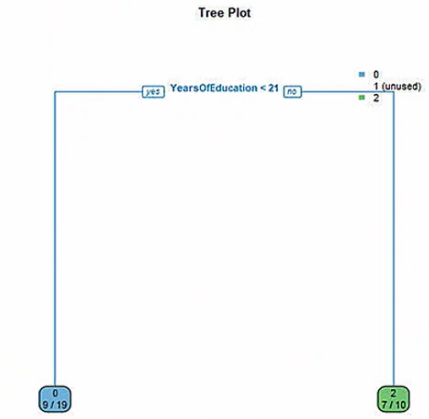
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.27778	0	1.00000	1.1111	0.13841
2	0.00000	1	0.72222	1.1111	0.13841

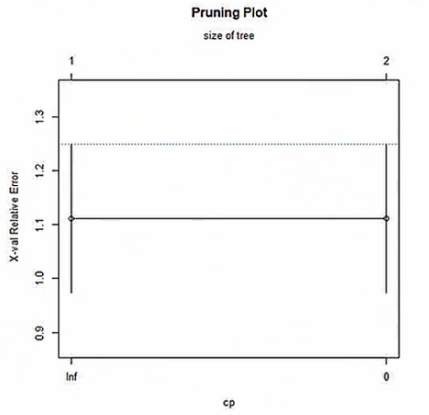
Leaf Summary

node), split, n, loss, yval, (yprob)
* denotes terminal node

```
1) root 29 18 0 (0.37931034 0.34482759 0.27586207)
  2) YearsOfEducation< 20.5 19 10 0 (0.47368421 0.47368421 0.05263158) *
  3) YearsOfEducation>=20.5 10  3 2 (0.20000000 0.10000000 0.70000000) *
```

Add these values and compare to the xError.
If sum > xError, then prune that branch.





BAGGING

FOREST MODEL

HISTORICAL DATA

ID	Predictor	Predictor	Predictor	Target
1	1	.7	5	124
2	3	.6	6	115
3	5	.5	5	101
4	3	.5	8	128
5	4	.4	6	111
6	5	.6	3	129

Rows are randomly sampled for inclusion in a given bag.

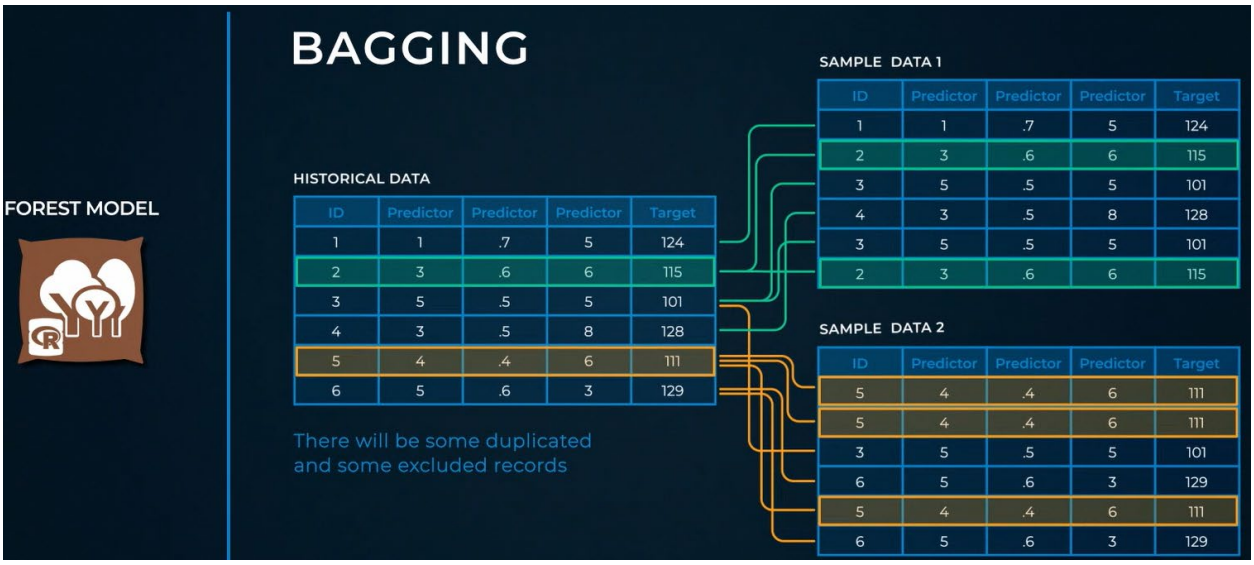
To make bags even, rows are duplicated

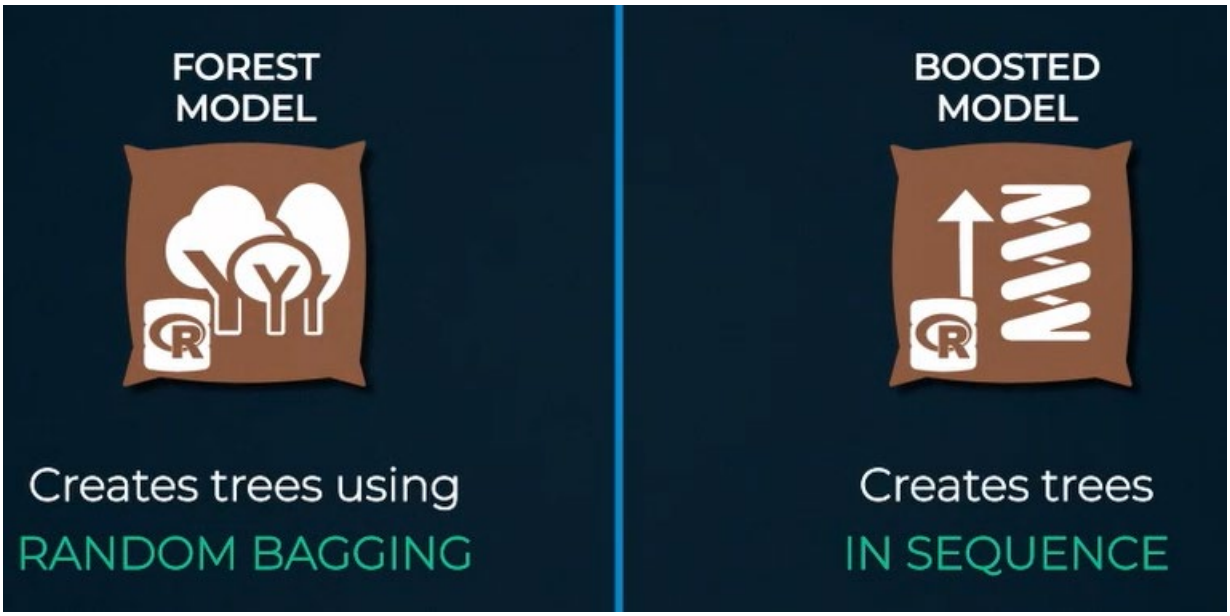
SAMPLE DATA 1

ID	Predictor	Predictor	Predictor	Target
1	1	.7	5	124
2	3	.6	6	115
3	5	.5	5	101
4	3	.5	8	128
3	5	.5	5	101
2	3	.6	6	115

SAMPLE DATA 2

ID	Predictor	Predictor	Predictor	Target
5	4	.4	6	111
5	4	.4	6	111
3	5	.5	5	101
6	5	.6	3	129
5	4	.4	6	111
6	5	.6	3	129

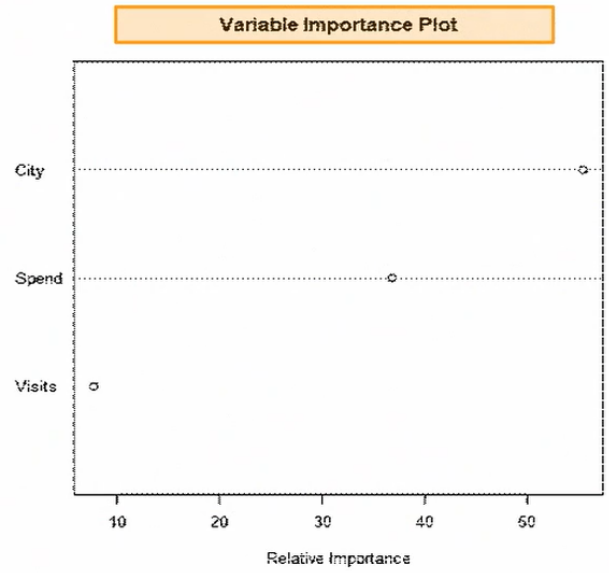

FOREST MODEL
BAGGING
HISTORICAL DATA
ID Predictor Predictor Predictor Target
1 1 .7 5 124
2 3 .6 6 115
3 5 .5 5 101
4 3 .5 8 128
5 4 .4 6 111
6 5 .6 3 129
There will be some duplicated and some excluded records
SAMPLE DATA 1
ID Predictor Predictor Predictor Target
1 1 .7 5 124
2 3 .6 6 115
3 5 .5 5 101
4 3 .5 8 128
3 5 .5 5 101
2 3 .6 6 115
SAMPLE DATA 2
ID Predictor Predictor Predictor Target
5 4 .4 6 111
5 4 .4 6 111
3 5 .5 5 101
6 5 .6 3 129
5 4 .4 6 111
6 5 .6 3 129

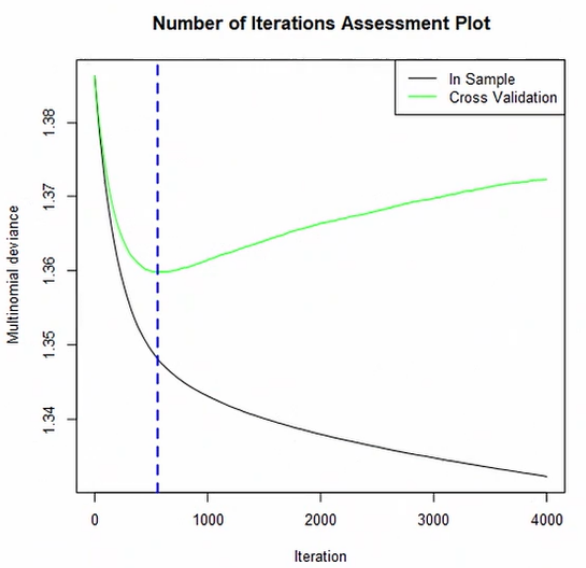

FOREST MODEL
Creates trees using
RANDOM BAGGING
BOOSTED MODEL
Creates trees
IN SEQUENCE

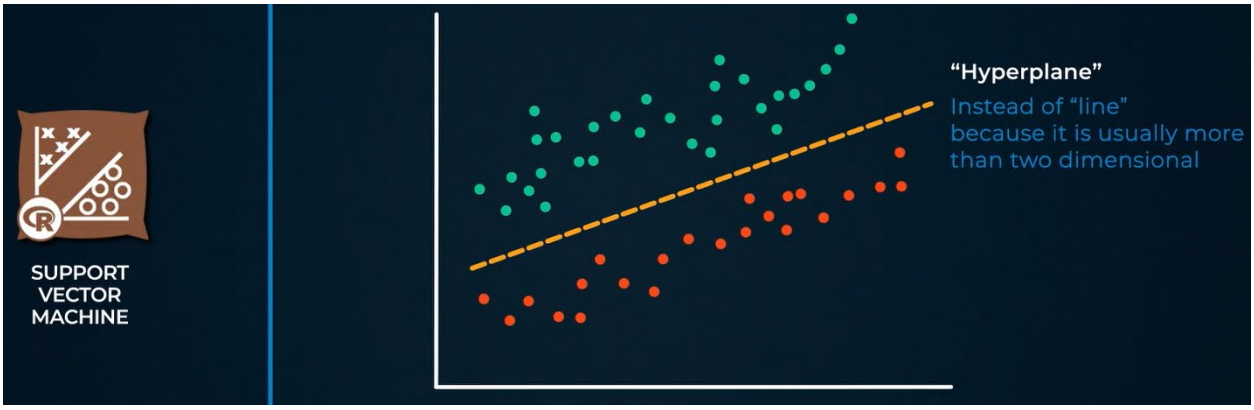
Report for Boosted Model Boosted

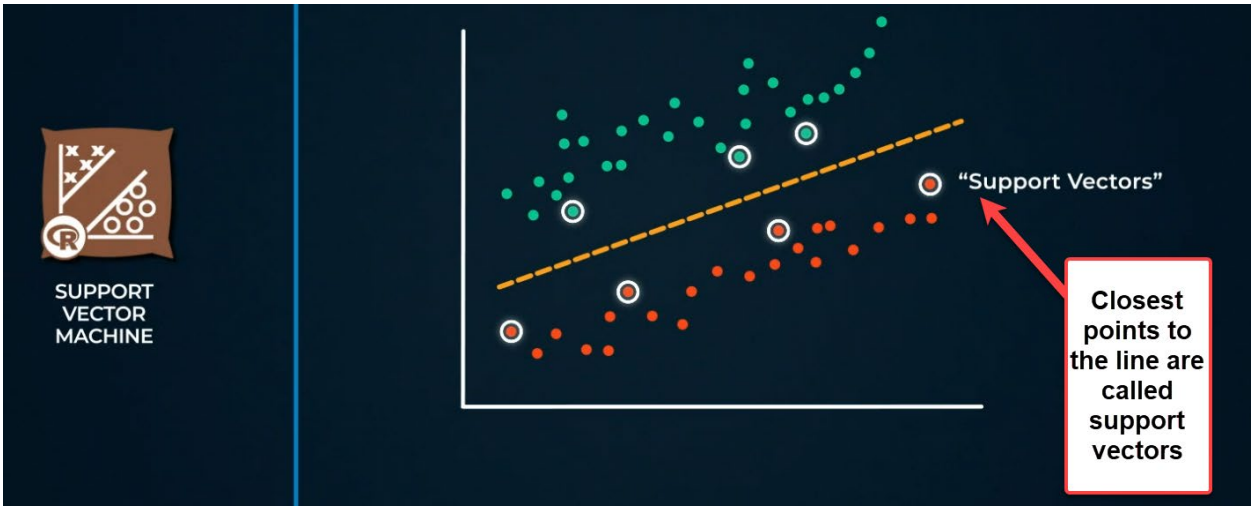
Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 557









Report for Support Vector Machine Model: Donate_SVM

Model Summary

Call: svm(formula = Donate ~ Degrees + First_Years + Last_Years + Undergraduate + First_School + Last_School + Faculty_Staff + Intercollegiate + Intramural + Other_Activities + Gender + Child + Parent + Spouse + Telephone + Mail + Personal + Combined + Log_Degrees + Log_Last_Years + Log_Telephone + Log_Mail + Log_Degrees_Sq + Log_Last_Years_Sq + Log_Telephone_Sq + Log_Mail_Sq + Last_Years_Cat, data = the.data, type = "C-classification", probability = TRUE)

Target: Donate

Predictors: Degrees, First_Years, Last_Years, Undergraduate, First_School, Last_School, Faculty_Staff, Intercollegiate, Intramural, Other_Activities, Gender, Child, Parent, Spouse, Telephone, Mail, Personal, Combined, Log_Degrees, Log_Last_Years, Log_Telephone, Log_Mail, Log_Degrees_Sq, Log_Last_Years_Sq, Log_Telephone_Sq, Log_Mail_Sq, Last_Years_Cat

Cost: 1

Gamma: 0.0169491525423729

Model Performance

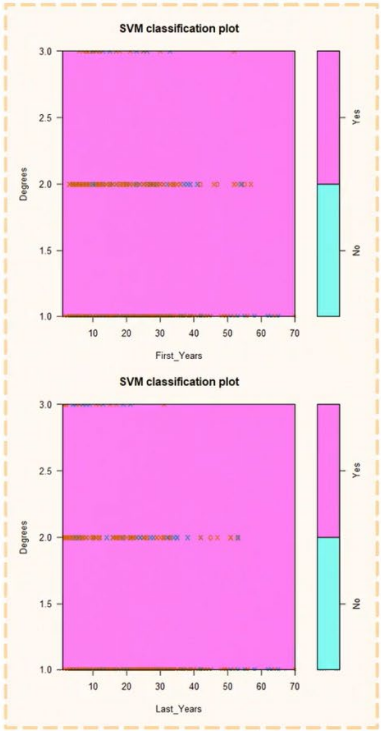
Confusion Matrix

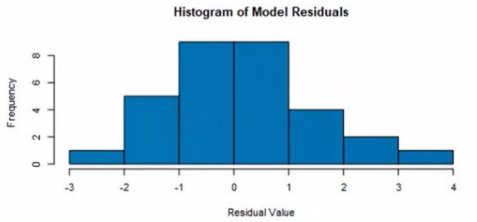
	No	Yes
No	430	189
Yes	173	408

Actuals are in columns, predicted in rows

***Note:** The performance here is solely based on training data set, thus good performance appears to be here does not always indicate a good models. It is possible that the model is overfit. The purpose of this "Model Performance" section is only for a quick reference.*

SVM Plots:





Report for Neural Network Model Neural Network

Basic Summary:

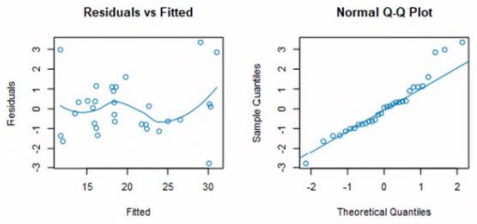
Call: nnet.formula(formula = MPG ~ Cyl + Displ + HP + Axle + Wght + QSec + VS + Trans + Gears + Carb, data = the.data, size = 10, linout = TRUE, rang = c(0.7), decay = 0.1, MaxNWts = 1000, maxit = 100)
Structure: A 10-10-1 network with 121 weights
Inputs: Cyl, Displ, HP, Axle, Wght, QSec, VS, Trans, Gears, Carb
Output(s): MPG
Options: Least-squares fitting , decay = 0.1
Final objective function value: Final objective function value: 84.11

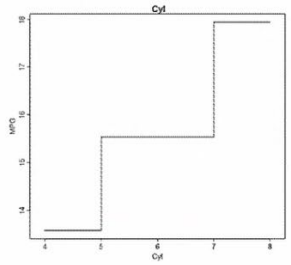
Input Layer

Hidden Layer

Output Layer

A confusion matrix (not shown here) will display for classification applications





Summary Report for Spline Model Spline

Call:
earth(formula = Egg_Pr ~ Beef_Pr + Cases + Cereal_Pr + Chicken_Pr + Easter + First_Week + Month + Pork_Pr, data = the.data, glm = list(family = gaussian), minspan = 0)

Coefficients:

Term	Value
(Intercept)	9.779e+01
h(Cases-108220)	-1.674e-04
h(108220-Cases)	2.420e-04
MonthDecember	6.853e+00
MonthMarch	6.091e+00
MonthFebruary	4.776e+00
MonthAugust	-1.075e+01
h(Pork_Pr-168.52)	5.086e-01
h(Cereal_Pr-119.89)	7.460e-01
MonthJune	-7.751e+00
MonthJuly	-6.127e+00
MonthMay	-5.397e+00
h(Pork_Pr-145.61)	-3.911e-01

Categorical Predictor

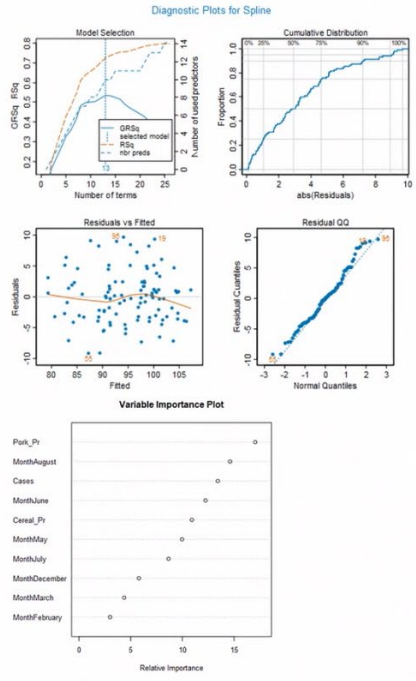
Selected 13 of 26 terms, and 10 of 20 predictors

Importance: Pork_Pr, MonthAugust, Cases, MonthJune, Cereal_Pr, MonthMay, MonthJuly, MonthDecember, MonthMarch, MonthFebruary

GCV 30.29 RSS 1818 GRSq 0.5323 RSq 0.7249

GLM null.deviance 6607 (103 dof) deviance 1818 (91 dof)

GLM Model McFadden R-Squared: 0.7249



Summary Report for Spline Model Spline

Call:

earth(formula = Egg_Pr ~ Beef_Pr + Cases + Cereal_Pr + Chicken_Pr + Easter + First_Week + Month + Pork_Pr, data = the.data, glm = list(family = gaussian), minspan = 0)

Coefficients:

	Term	Value
Continuous Predictors	(Intercept)	9.779e+01
	h(Cases-108220)	-1.674e-04
	h(108220-Cases)	2.420e-04
	MonthDecember	6.853e+00
	MonthMarch	6.091e+00
	MonthFebruary	4.776e+00
	MonthAugust	-1.075e+01
	h(Pork_Pr-168.52)	5.086e-01
	h(Cereal_Pr-119.89)	7.460e-01
	MonthJune	-7.751e+00
	MonthJuly	-6.127e+00
	MonthMay	-5.397e+00
	h(Pork_Pr-145.61)	-3.911e-01

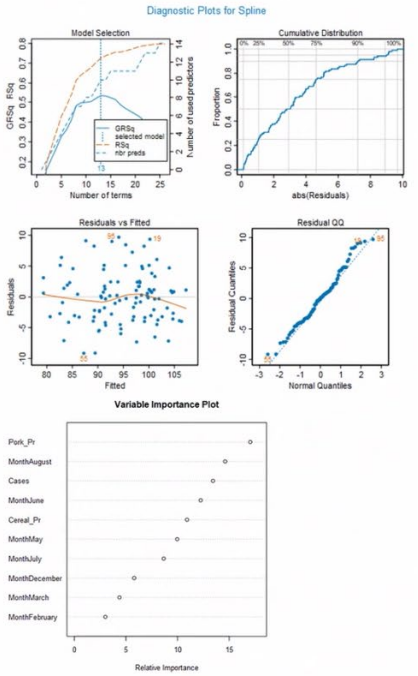
Selected 13 of 26 terms, and 10 of 20 predictors

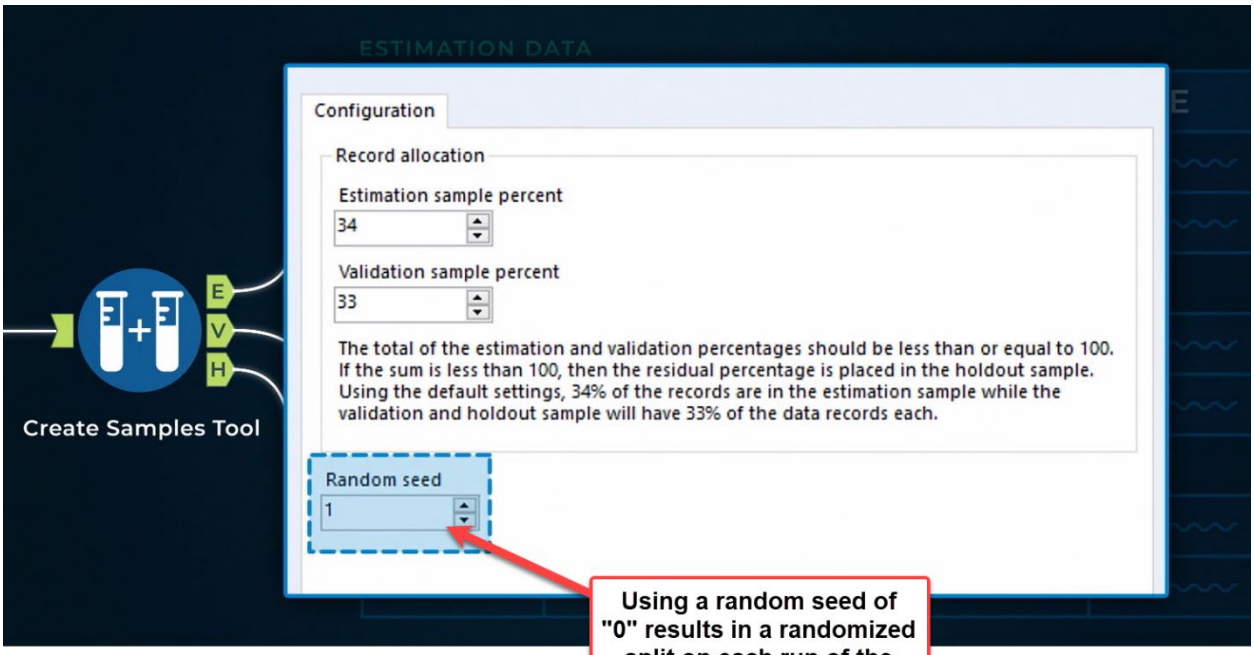
Importance: Pork_Pr, MonthAugust, Cases, MonthJune, Cereal_Pr, MonthMay, MonthJuly, MonthDecember, MonthMarch, MonthFebruary

GCV 30.29 RSS 1818 GRSq 0.5323 RSq 0.7249

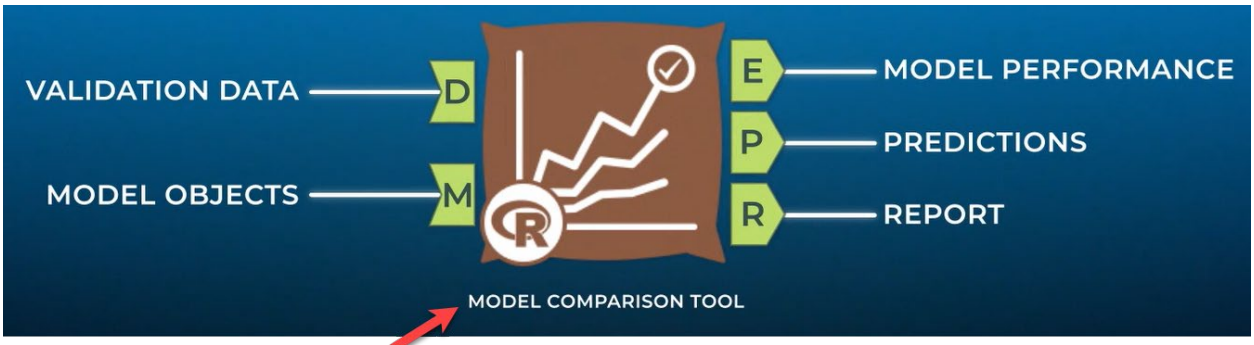
GLM null.deviance 6607 (103 dof) deviance 1818 (91 dof)

GLM Model McFadden R-Squared: 0.7249





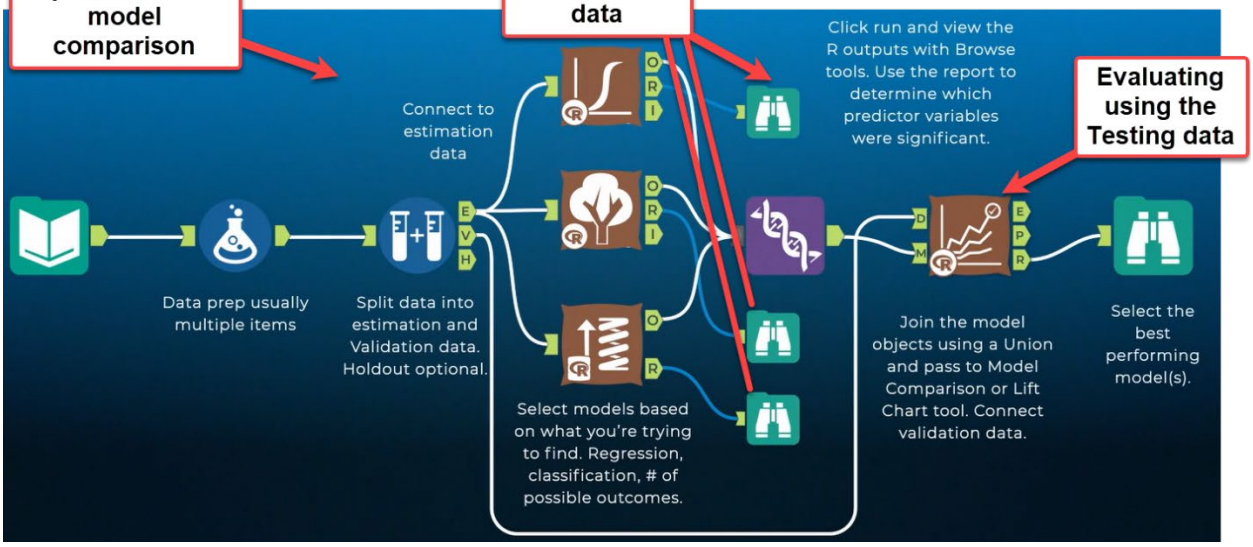
Using a random seed of "0" results in a randomized split on each run of the workflow, which is generally *not* recommended



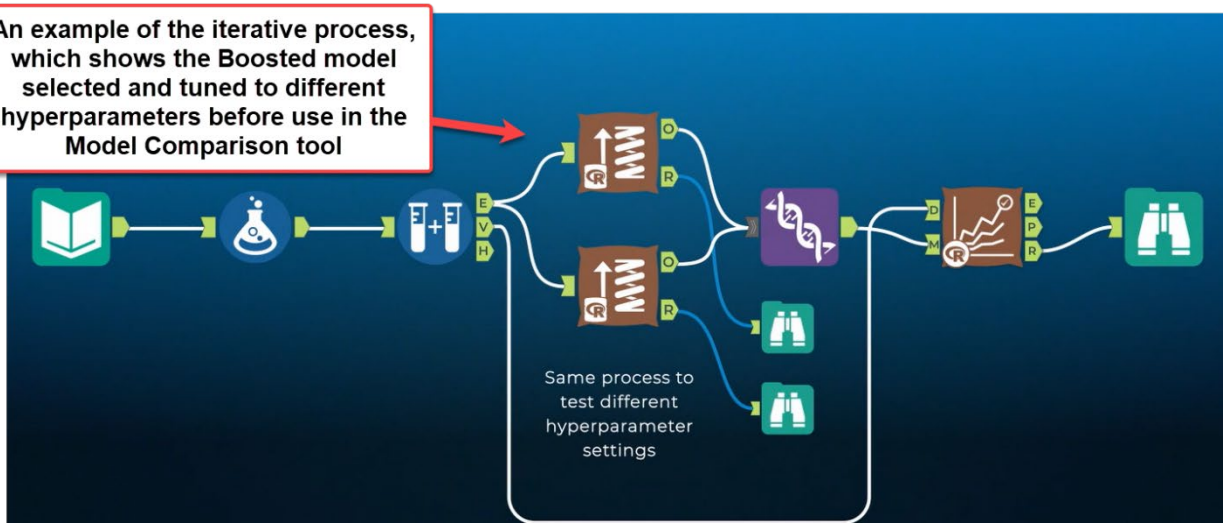
Requires no configuration

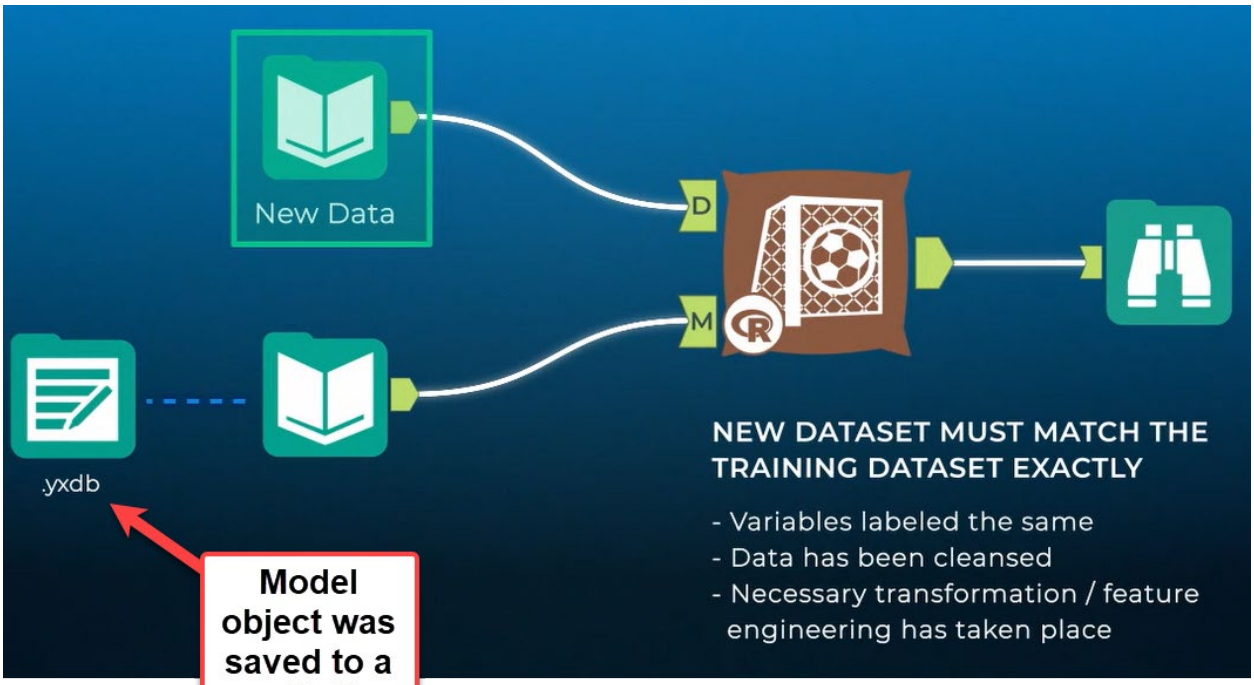
Example design pattern for a model comparison

Reporting on the Training data



An example of the iterative process, which shows the Boosted model selected and tuned to different hyperparameters before use in the Model Comparison tool




New Data
D
M
.yxdb
NEW DATASET MUST MATCH THE TRAINING DATASET EXACTLY
- Variables labeled the same
- Data has been cleansed
- Necessary transformation / feature engineering has taken place
Model object was saved to a yxdb file

Question 1

Which of the following algorithms can be used for classification. Select all that apply.

- [x] Support Vector Machine
- [x] Boosted Model
- [x] Neural Network
- [x] Logistic Regression
- [ ] Gamma Regression

Question 2

Why would you split data into estimation and validation data?

- ○ None of these
- ○ To stratify the datasets, ensuring equal representation of target classes
- ○ To ensure the training & testing datasets have equal number of records
- ● To ensure the model is not overfit to the training data
- ○ This step is not necessary

Question 3

A higher AIC indicates a better model.

- True
- False

Question 4

What is bagging?

- Separating data into estimation and validation datasets
- Randomly dividing records with replacement to create different datasets
- Sampling records a model did not perform well in order to train the next model
- None of these

Question 5

What is the main advantage to outputting your model object as a .yxdb?

- [x] Saves time by not having to retrain the model
- [ ] You can use the model object in other workflows
- [ ] You can upload the model to your gallery
- [ ] None of these

Question 6

Which of the following algorithms is most sensitive to outliers?

- Decision Tree
- Support Vector Machine
- Depends on the dataset
- Linear Regression

Question 7

How are the reports from the Model Comparison tool different from the reports built into model algorithms? Select all that apply.

- [x] They represent performance on different datasets
- [ ] One contains a confusion matrix and the other does not
- [ ] The Model Comparison report is better for identifying the most important predictor variables
- [ ] There is no difference between the reports
- [ ] None of these

Remember that the data sources attached to each of these should be different, but that the objective of each report is also different. Algorithm reports focus on what attributes contributed to the model's performance so you can improve that algorithm, while the Model Comparison is used to evaluate which algorithm fits the current dataset best.

Question 8

What does the *Random Seed* configuration on the Create Samples tool control?

- ○ It controls the percentage of records that become holdout data
- ○ It specifies the 1 in N chance that a record is included
- ● It changes which rows appear in the estimation and validation datasets
- ○ None of these

Question 9

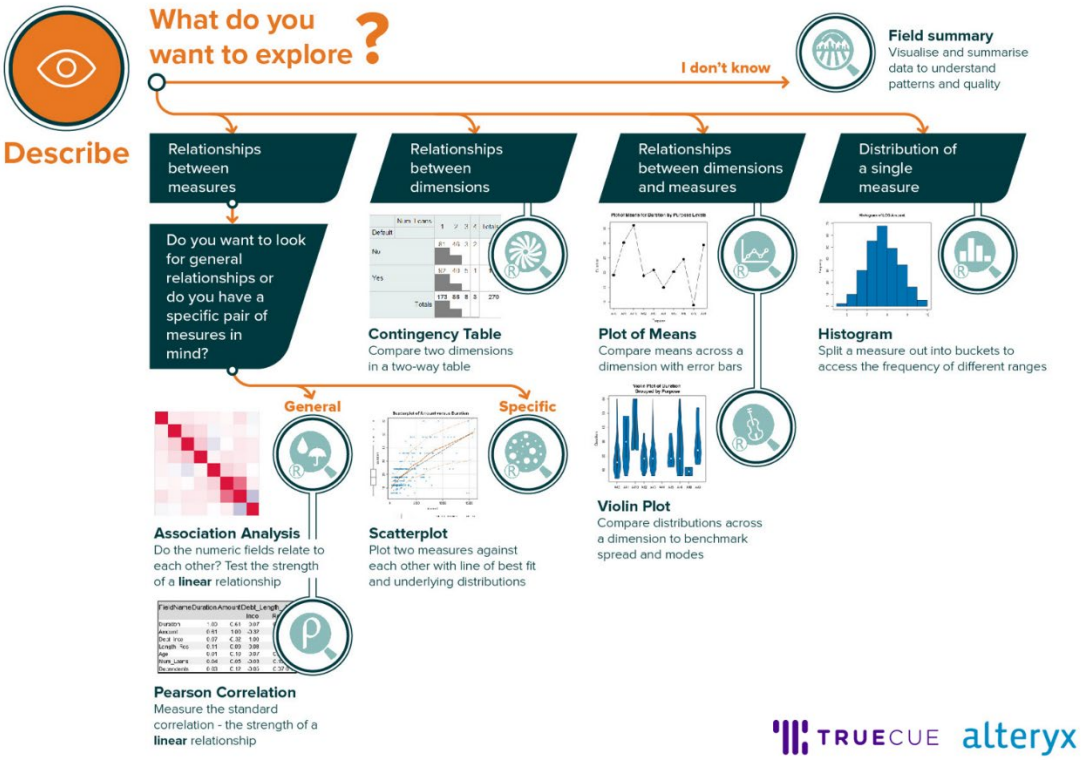
Which factors are important when deciding which modeling algorithm to use? Select all that apply.

- [x] The number of possible classifications
- [x] Your ability to explain the results of the model
- [x] The datatype of the target variable
- [x] The distribution of the target variable

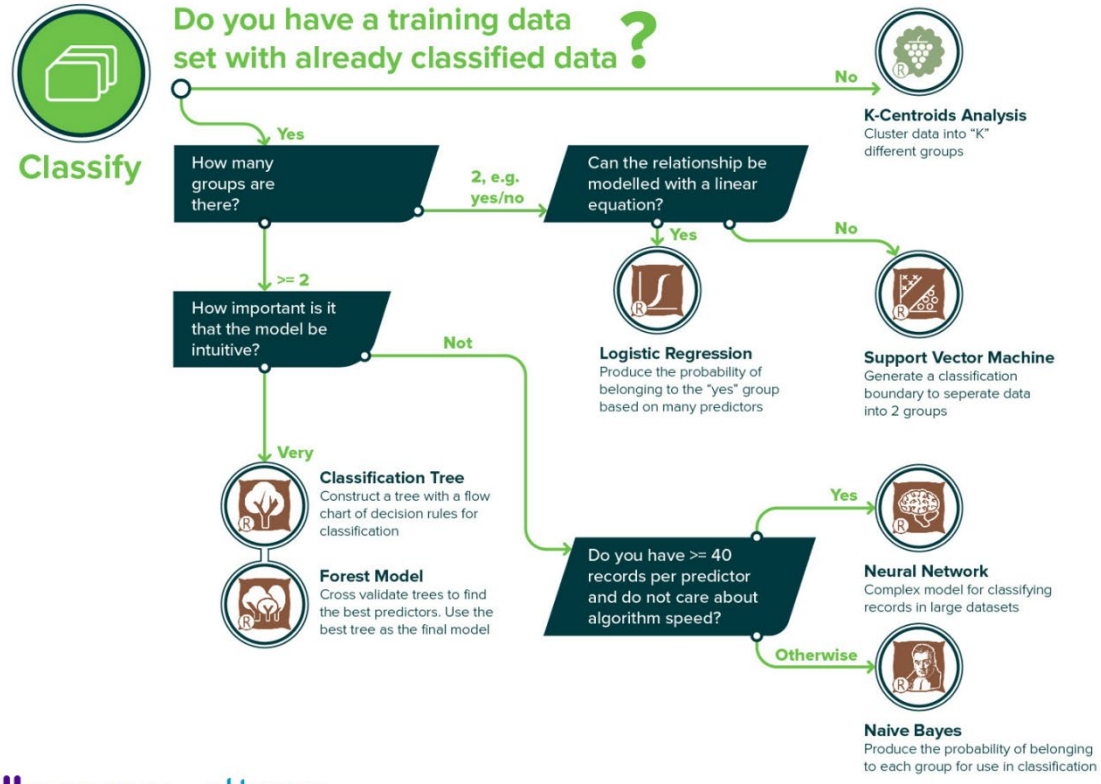
Question 10

Where is the best place to find which variables are important for a given model?

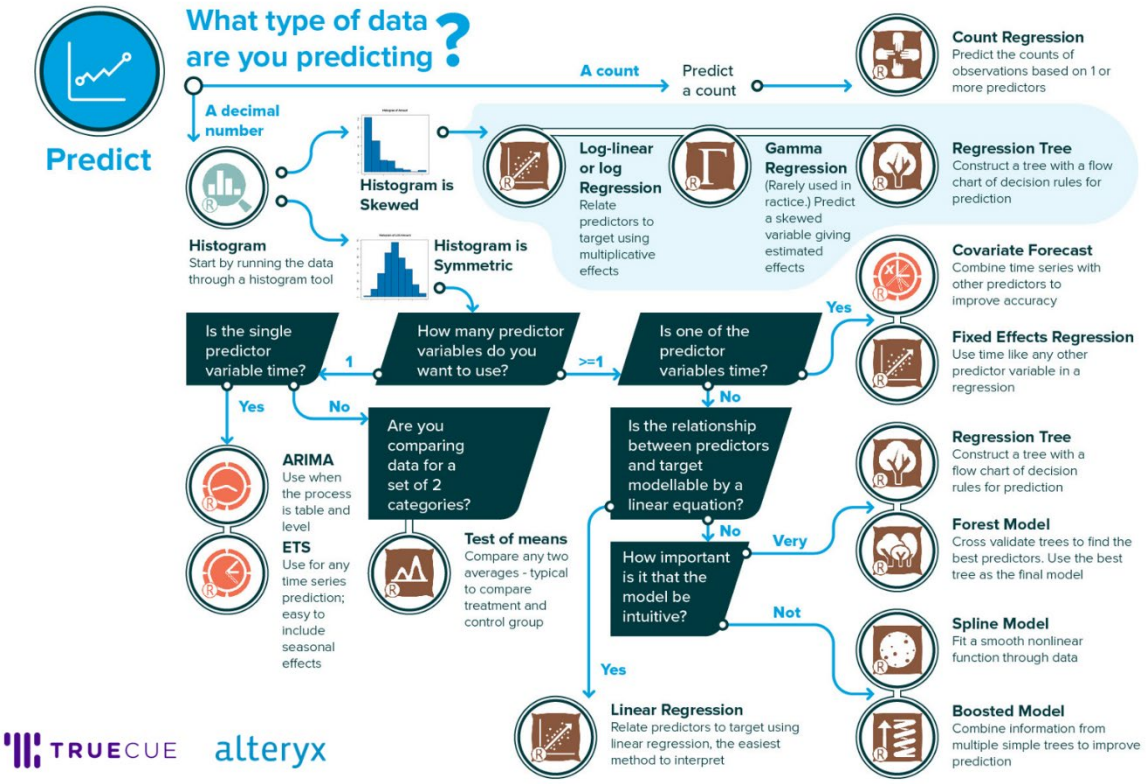
- The Score tool
- The Model Comparison Report
- The model algorithm's Report anchor
- The Field Summary tool

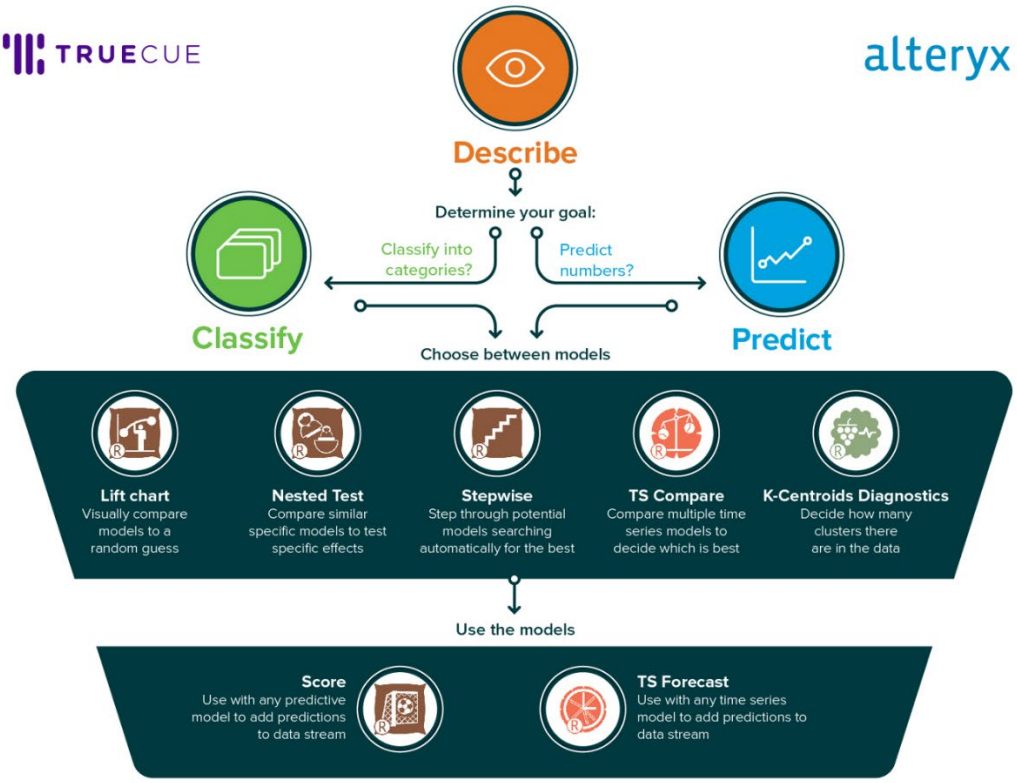


TRUECUE alteryx



TRUECUE alteryx


Predict
What type of data are you predicting?
A count
Predict a count
Count Regression
Predict the counts of observations based on 1 or more predictors
A decimal number
Histogram
Start by running the data through a histogram tool
Histogram is Skewed
Histogram is Symmetric
Log-linear or log Regression
Relate predictors to target using multiplicative effects
Gamma Regression
(Rarely used in ractice.) Predict a skewed variable giving estimated effects
Regression Tree
Construct a tree with a flow chart of decision rules for prediction
Covariate Forecast
Combine time series with other predictors to improve accuracy
Fixed Effects Regression
Use time like any other predictor variable in a regression
Is the single predictor variable time?
1
How many predictor variables do you want to use?
>=1
Is one of the predictor variables time?
Yes
No
Yes
No
ARIMA
Use when the process is table and level
ETS
Use for any time series prediction; easy to include seasonal effects
Are you comparing data for a set of 2 categories?
Test of means
Compare any two averages - typical to compare treatment and control group
Is the relationship between predictors and target modellable by a linear equation?
No
Very
Not
How important is it that the model be intuitive?
Yes
Linear Regression
Relate predictors to target using linear regression, the easiest method to interpret
Regression Tree
Construct a tree with a flow chart of decision rules for prediction
Forest Model
Cross validate trees to find the best predictors. Use the best tree as the final model
Spline Model
Fit a smooth nonlinear function through data
Boosted Model
Combine information from multiple simple trees to improve prediction
TRUECUE
alteryx


TRUECUE
alteryx
Describe
Determine your goal:
Classify into categories?
Predict numbers?
Classify
Predict
Choose between models
Lift chart
Visually compare models to a random guess
Nested Test
Compare similar specific models to test specific effects
Stepwise
Step through potential models searching automatically for the best
TS Compare
Compare multiple time series models to decide which is best
K-Centroids Diagnostics
Decide how many clusters there are in the data
Use the models
Score
Use with any predictive model to add predictions to data stream
TS Forecast
Use with any time series model to add predictions to data stream

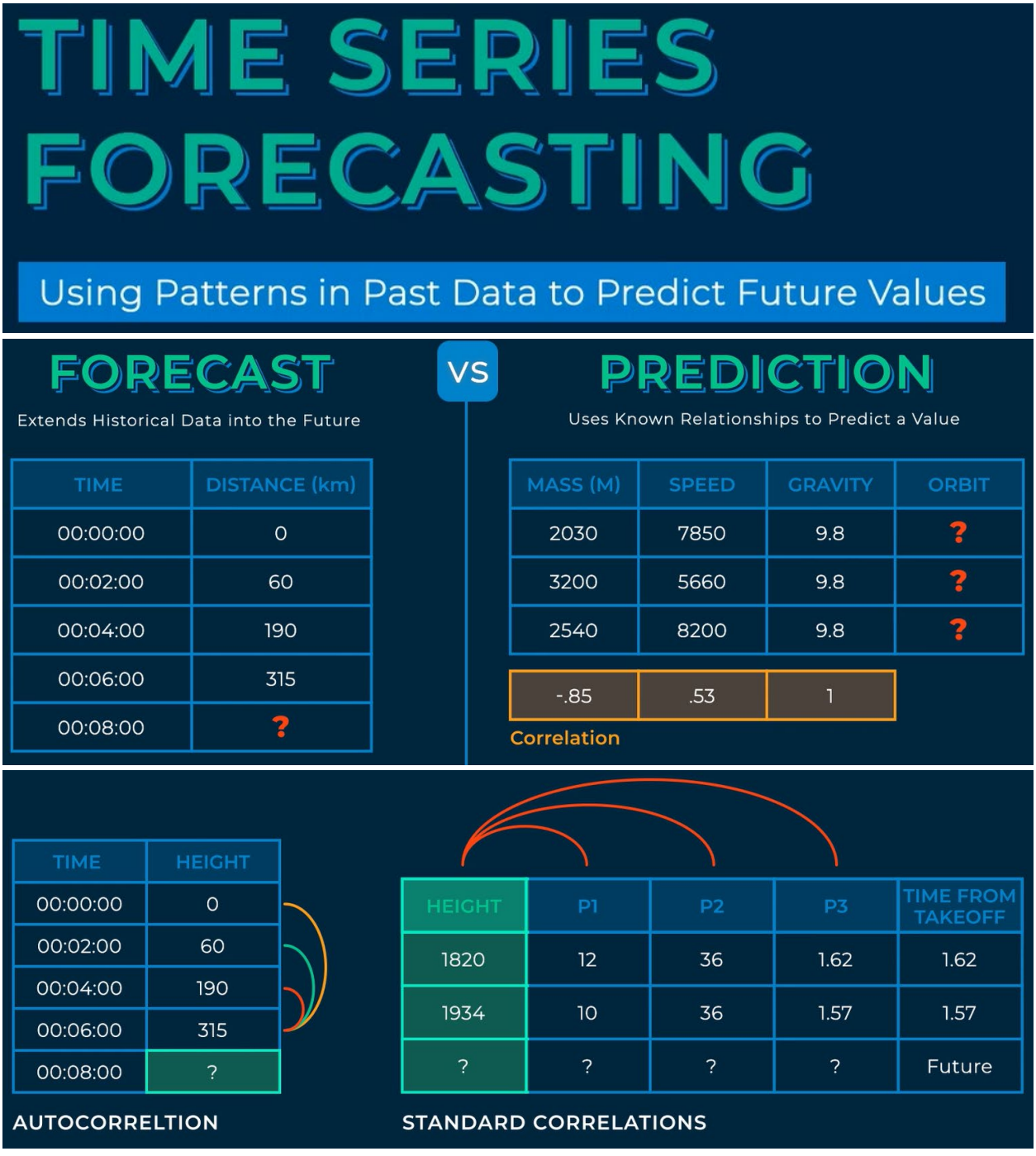
Record	Report
1	**Summary Report for Decision Tree Model IrisDecisionTree**
2	Call: rpart(formula = Species ~ SepalLengthCm + SepalWidthCm + PetalLengthCm + PetalWidthCm, data = the.data, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 0, usesurrogate = 0, surrogatestyle = 0)
3	Model Summary Variables actually used in tree construction: [1] PetalLengthCm PetalWidthCm Root node error: 100/150 = 0.66667 n= 150
4	*Pruning Table*

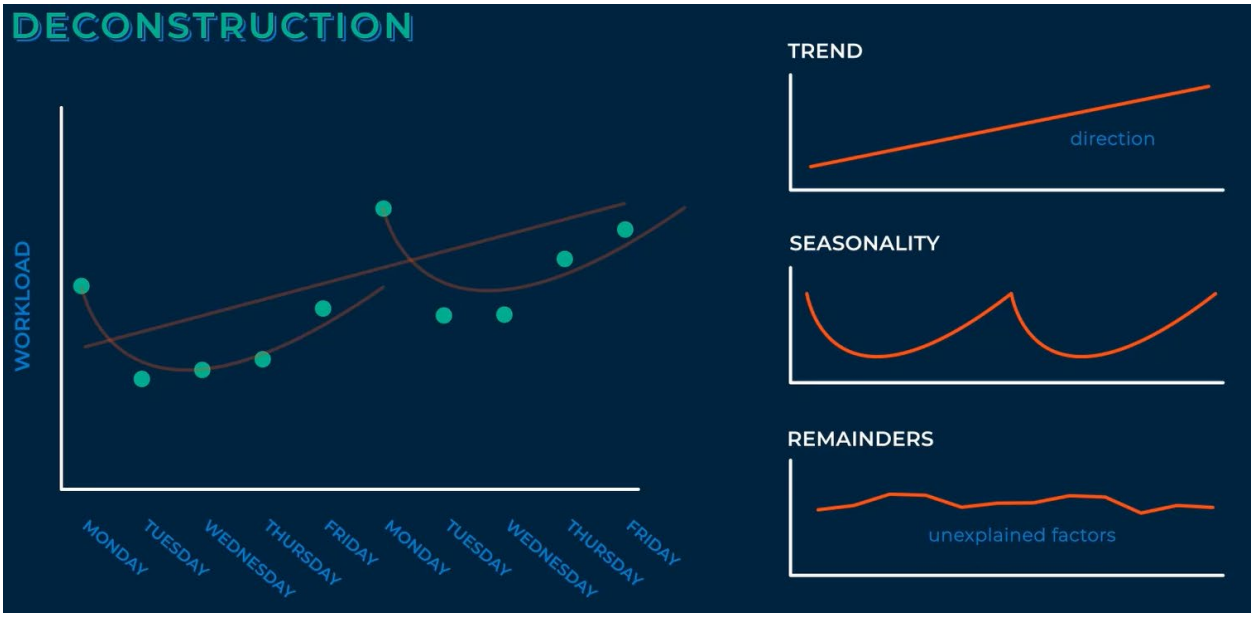
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.50	0	1.00	1.17	0.050735
2	0.44	1	0.50	0.73	0.061215
3	0.00	2	0.06	0.10	0.030551

Record	Report
6	Leaf Summary node), split, n, loss, yval, (yprob) * denotes terminal node 1) root 150 100 Iris-setosa (0.33333333 0.33333333 0.33333333) 2) PetalLengthCm< 2.45 50 0 Iris-setosa (1.00000000 0.00000000 0.00000000) * 3) PetalLengthCm>=2.45 100 50 Iris-versicolor (0.00000000 0.50000000 0.50000000) 6) PetalWidthCm< 1.75 54 5 Iris-versicolor (0.00000000 0.90740741 0.09259259) * 7) PetalWidthCm>=1.75 46 1 Iris-virginica (0.00000000 0.02173913 0.97826087) *

Regarding the *Decision Tree* tool: Classification Trees are typically evaluated with confusion matrices and F1-Scores, whereas Regression Trees are assessed with R^2 and Mean Square Error (MSE). A rule of thumb is to select the lowest level where rel_error + xstd < xerror.

What is Time Series Forecasting?





REGULARLY SPACED INTERVALS

DATE	QUANTITY	UNIT
2014-05-10	1372	kg
2015-05-10	1406	kg
2016-05-10	1525	kg
2017-06-10	1634	kg
2018-05-10	1892	kg
2019-05-10	1750	lb
2020-05-10	2039	lb

Irregular Month

Inconsistent Units

Consistent space and measurement, as well as addressing any missing values, between each observation are required in Time Series Analysis

DATA REQUIREMENTS

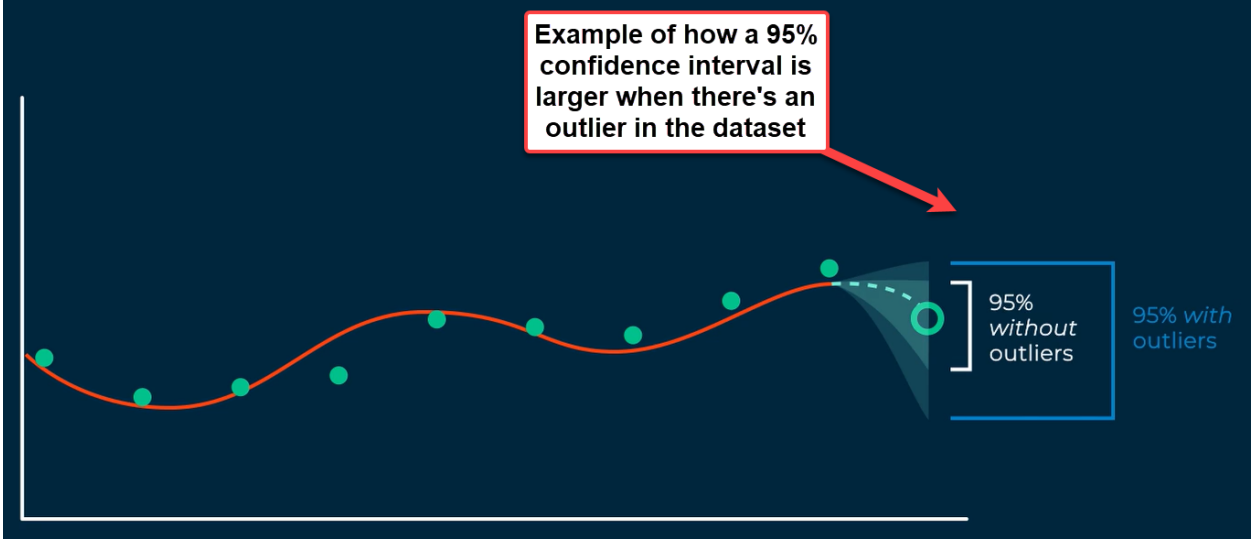
An example of Time Series Analysis, with the target variable being "Quantity."

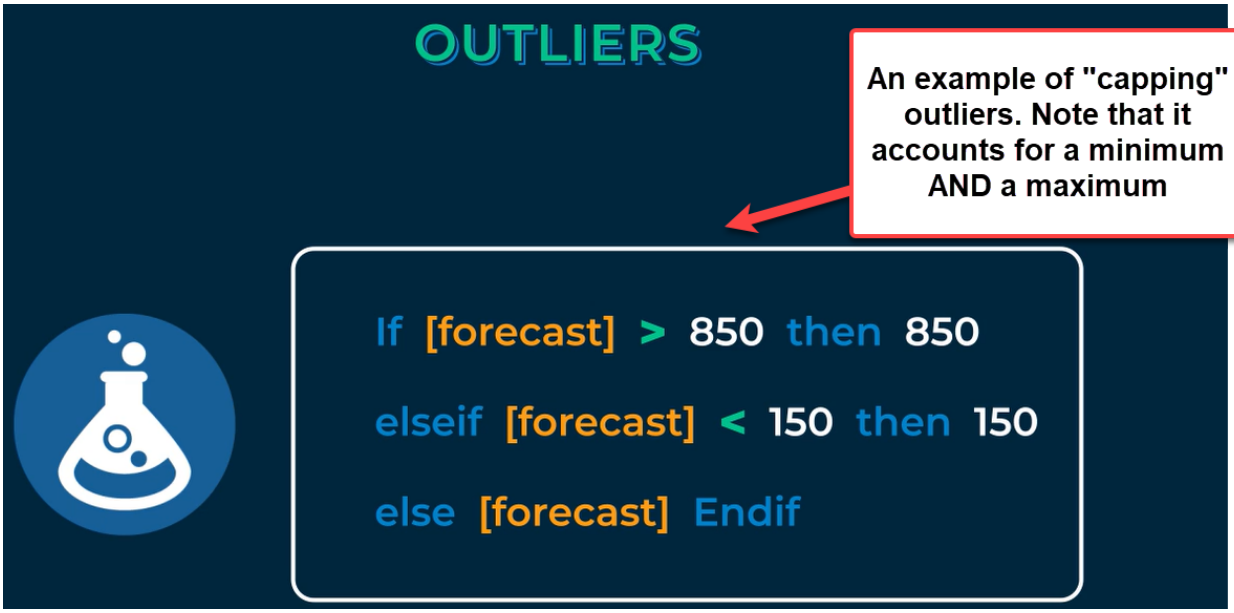
DATE	QUANTITY	UNIT
2014-05-10	1372	kg
2015-05-10	1406	kg
2016-05-10	1525	kg
2017-05-10	1634	kg
2018-05-10	1892	kg
2019-05-10	1750	lb
2020-05-10	2039	lb

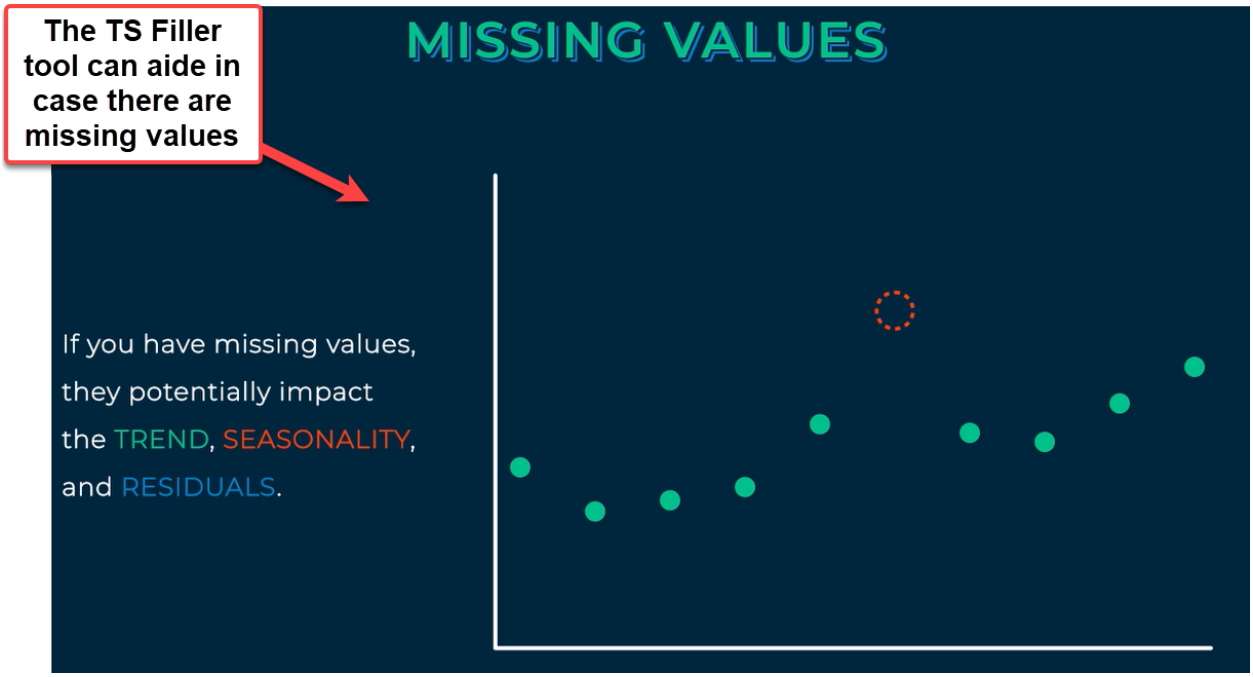
target variable

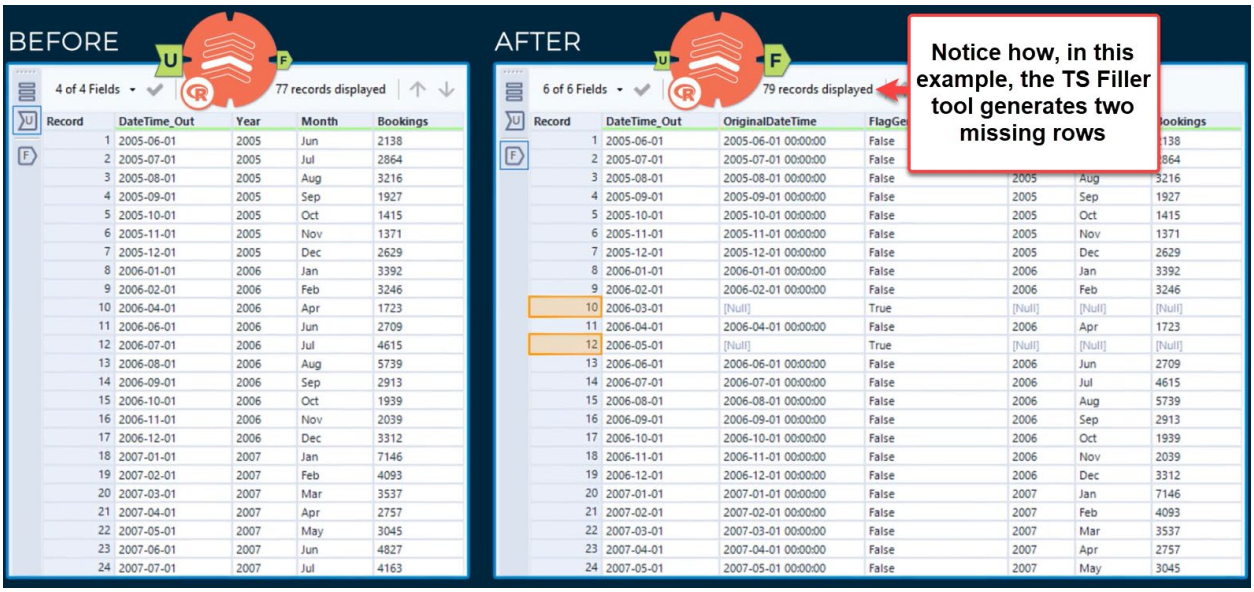
- Must be Numeric
- Single Column
- Consistent Units
- Date = Ascending Order

OUTLIERS




OUTLIERS
An example of "capping" outliers. Note that it accounts for a minimum AND a maximum
If [forecast] > 850 then 850
elseif [forecast] < 150 then 150
else [forecast] Endif


The TS Filler tool can aide in case there are missing values
MISSING VALUES
If you have missing values, they potentially impact the TREND, SEASONALITY, and RESIDUALS.



Record	DateTime_Out	Year	Month	Bookings
1	2005-06-01	2005	Jun	2138
2	2005-07-01	2005	Jul	2864
3	2005-08-01	2005	Aug	3216
4	2005-09-01	2005	Sep	1927
5	2005-10-01	2005	Oct	1415
6	2005-11-01	2005	Nov	1371
7	2005-12-01	2005	Dec	2629
8	2006-01-01	2006	Jan	3392
9	2006-02-01	2006	Feb	3246
10	2006-04-01	2006	Apr	1723
11	2006-06-01	2006	Jun	2709
12	2006-07-01	2006	Jul	4615
13	2006-08-01	2006	Aug	5739
14	2006-09-01	2006	Sep	2913
15	2006-10-01	2006	Oct	1939
16	2006-11-01	2006	Nov	2039
17	2006-12-01	2006	Dec	3312
18	2007-01-01	2007	Jan	7146
19	2007-02-01	2007	Feb	4093
20	2007-03-01	2007	Mar	3537
21	2007-04-01	2007	Apr	2757
22	2007-05-01	2007	May	3045
23	2007-06-01	2007	Jun	4827
24	2007-07-01	2007	Jul	4163

Record	DateTime_Out	OriginalDateTime	FlagGe			Bookings
1	2005-06-01	2005-06-01 00:00:00	False			2138
2	2005-07-01	2005-07-01 00:00:00	False			2864
3	2005-08-01	2005-08-01 00:00:00	False	2005	Aug	3216
4	2005-09-01	2005-09-01 00:00:00	False	2005	Sep	1927
5	2005-10-01	2005-10-01 00:00:00	False	2005	Oct	1415
6	2005-11-01	2005-11-01 00:00:00	False	2005	Nov	1371
7	2005-12-01	2005-12-01 00:00:00	False	2005	Dec	2629
8	2006-01-01	2006-01-01 00:00:00	False	2006	Jan	3392
9	2006-02-01	2006-02-01 00:00:00	False	2006	Feb	3246
10	2006-03-01	[Null]	True	[Null]	[Null]	[Null]
11	2006-04-01	2006-04-01 00:00:00	False	2006	Apr	1723
12	2006-05-01	[Null]	True	[Null]	[Null]	[Null]
13	2006-06-01	2006-06-01 00:00:00	False	2006	Jun	2709
14	2006-07-01	2006-07-01 00:00:00	False	2006	Jul	4615
15	2006-08-01	2006-08-01 00:00:00	False	2006	Aug	5739
16	2006-09-01	2006-09-01 00:00:00	False	2006	Sep	2913
17	2006-10-01	2006-10-01 00:00:00	False	2006	Oct	1939
18	2006-11-01	2006-11-01 00:00:00	False	2006	Nov	2039
19	2006-12-01	2006-12-01 00:00:00	False	2006	Dec	3312
20	2007-01-01	2007-01-01 00:00:00	False	2007	Jan	7146
21	2007-02-01	2007-02-01 00:00:00	False	2007	Feb	4093
22	2007-03-01	2007-03-01 00:00:00	False	2007	Mar	3537
23	2007-04-01	2007-04-01 00:00:00	False	2007	Apr	2757
24	2007-05-01	2007-05-01 00:00:00	False	2007	May	3045

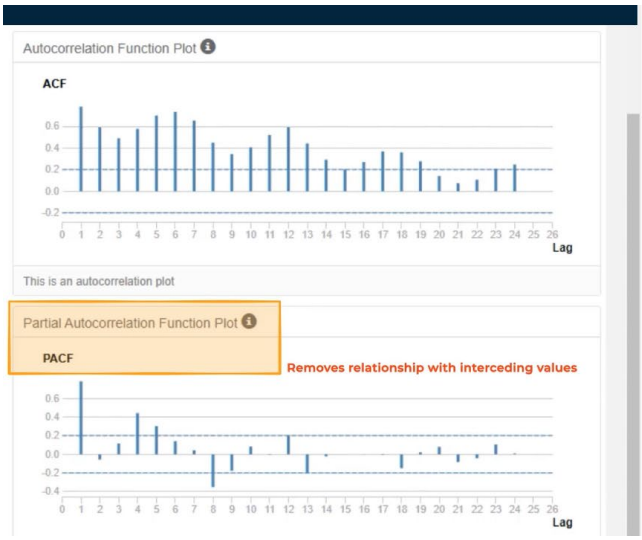
MOVING AVERAGE

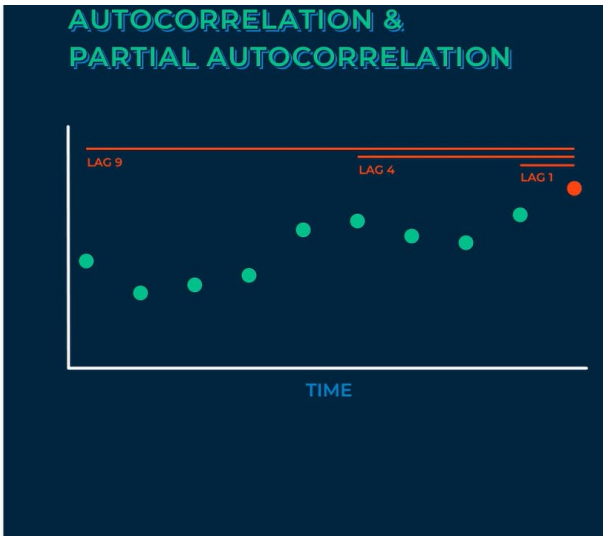
DATE	QUANTITY
2014-05-10	1372
2015-05-10	1406
2016-05-10	[null]
2017-05-10	1634
2018-05-10	1892
2019-05-10	1750
2020-05-10	2039

A common way to fill in a missing value is to use the moving average

$$= \frac{1406 + 1634}{2} = 1520$$

```
if IsNull([Y])
then (([Row-1:Y] + [Row+1:Y]) / 2)
else [Y] endif
```





Question 1

What are acceptable methods when accounting for outliers in Time Series models? Select all that apply.

- [ ] Reduce the size of the confidence intervals
- [x] Capping the values
- [ ] There is no need to account for outliers
- [x] Replace with the Moving Average
- [ ] They only effect ARIMA models

Question 2

Seasonality can refer to patterns in: (Select all that apply)

- [x] Years
- [x] Months
- [x] Days
- [x] Weeks
- [x] Quarters

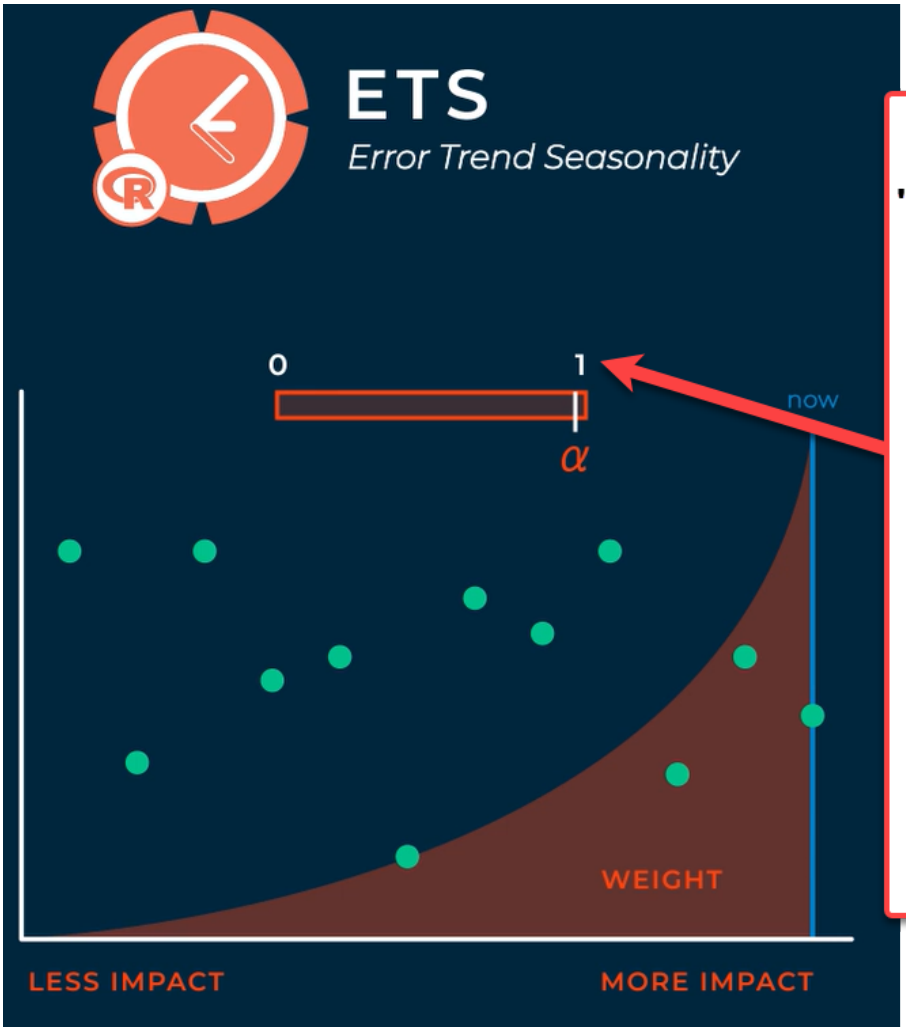
Question 3

Based on the dataset pictured, which type of prediction is possible? Select all that apply.

- [x] A prediction for the next week
- [x] A prediction for the next month
- [ ] A prediction for the next day

	Year	Week_Of	Total Consumption EUR	Total Price EUR	Total Consumption SIB	Total Price SIB
1	2006	2006-09-01 00:00:00	11460707	58252.76	3146358	58252.76
2	2006	2006-09-08 00:00:00	11882622	78173.55	3171588	78173.55
3	2006	2006-09-15 00:00:00	12333684	80129.9	3178116	80129.9
4	2006	2006-09-22 00:00:00	12485116	80867.01	3395715	80867.01
5	2006	2006-09-29 00:00:00	12743314	72914.86	3478068	72914.86
6	2006	2006-10-06 00:00:00	13216430	78234.16	3697051	78234.16
7	2006	2006-10-13 00:00:00	13881852	76459.17	3770579	76459.17
8	2006	2006-10-20 00:00:00	14317821	80313.18	3812568	80313.18
9	2006	2006-10-27 00:00:00	14216725	70492.8	3790422	70492.8
10	2006	2006-11-03 00:00:00	14421793	69554	3780349	69554
11	2006	2006-11-10 00:00:00	14863115	73054.56	3880309	73054.56
12	2006	2006-11-17 00:00:00	15169597	74299.83	3995643	74299.83
13	2006	2006-11-24 00:00:00	15326853	80688.2	4249540	80688.2

ETS and ARIMA



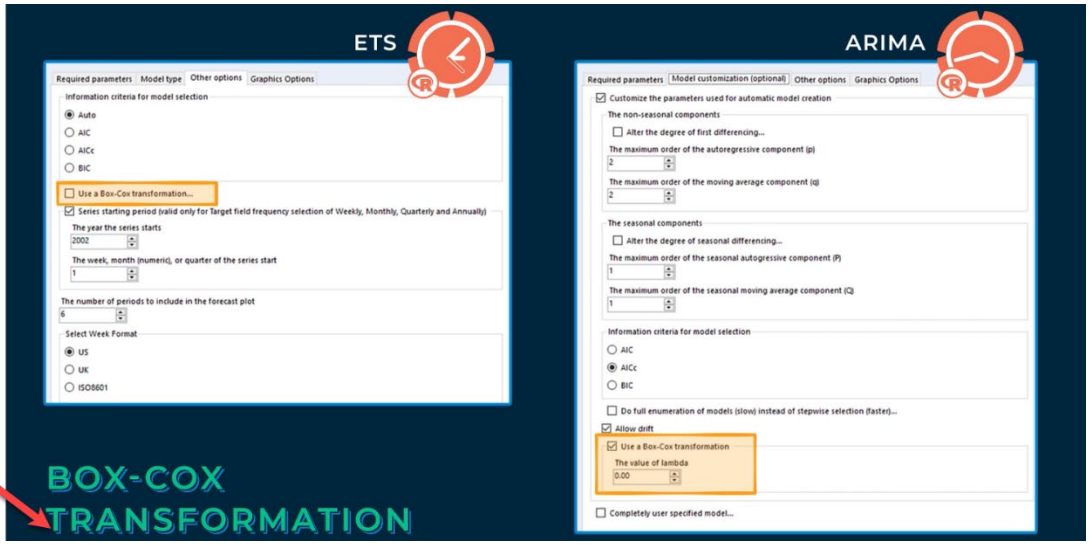
For ETS: An "alpha" value that is closer to 1 means the algorithm gives exponentially more weight to the most recent observations when compared to older observations. When "alpha" is 0, all data points are weighted equally.

STATIONARITY

- Many time series algorithms assume a dataset's mean and variance are stable over time.

- Stationary data reduces forecast uncertainty.

- Most real world data is non-stationary and will require mathematical transformation.



Used to stabilize the variance of a dataset

BOX-COX TRANSFORMATION

Has little effect on point forecasts,
but affects the size of confidence intervals.

ETS: These parameters are automatically assigned by the algorithm

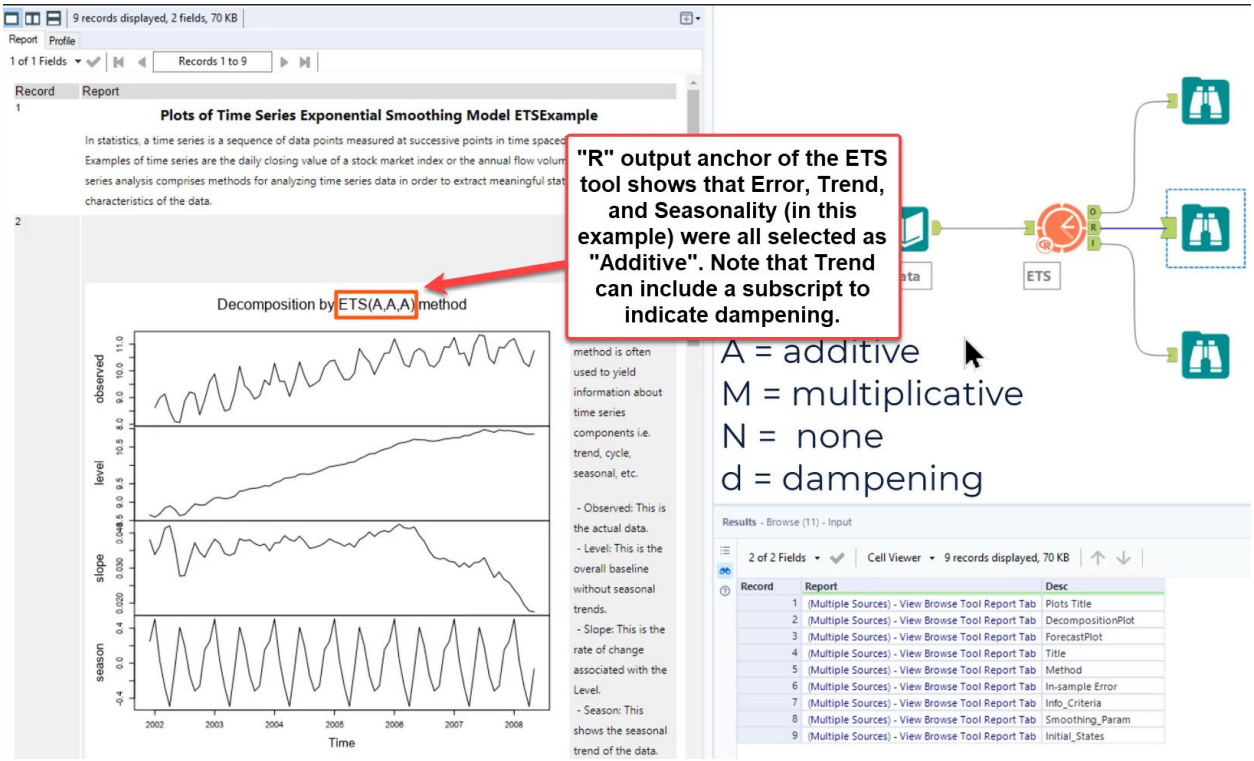
SMOOTHING PARAMETER

α = LEVEL

β = TREND

γ = SEASON

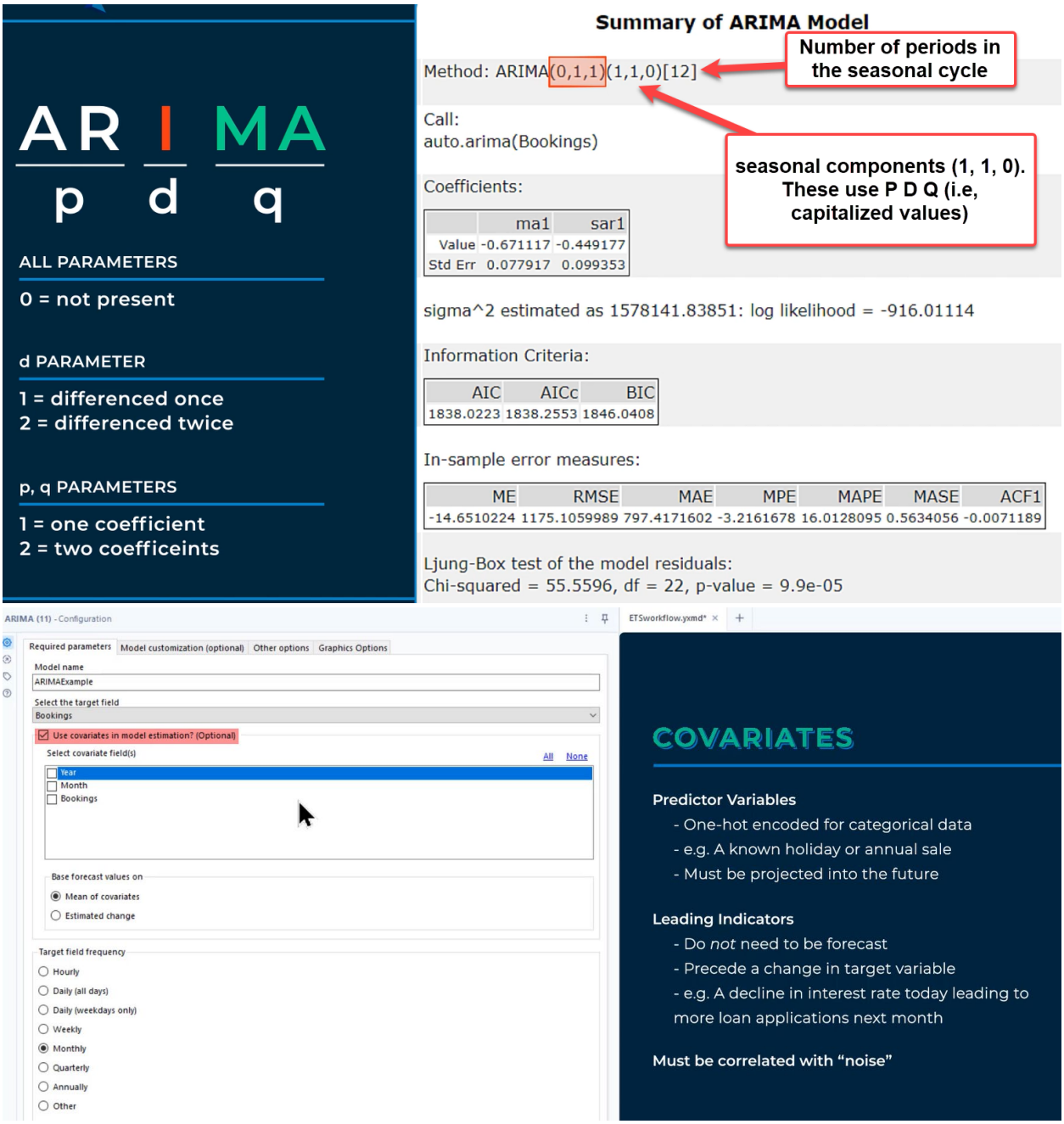
ϕ = DAMPENING

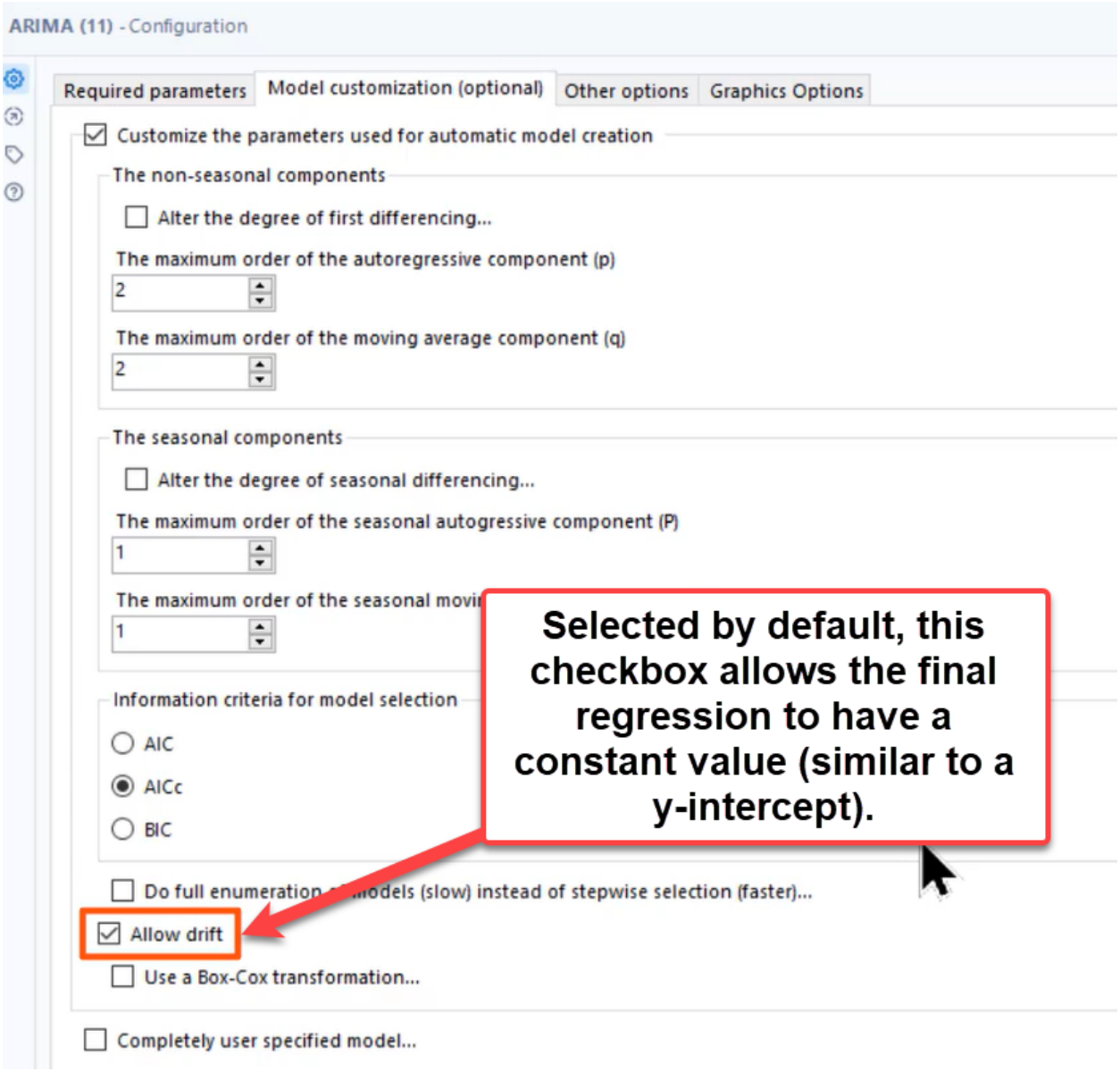


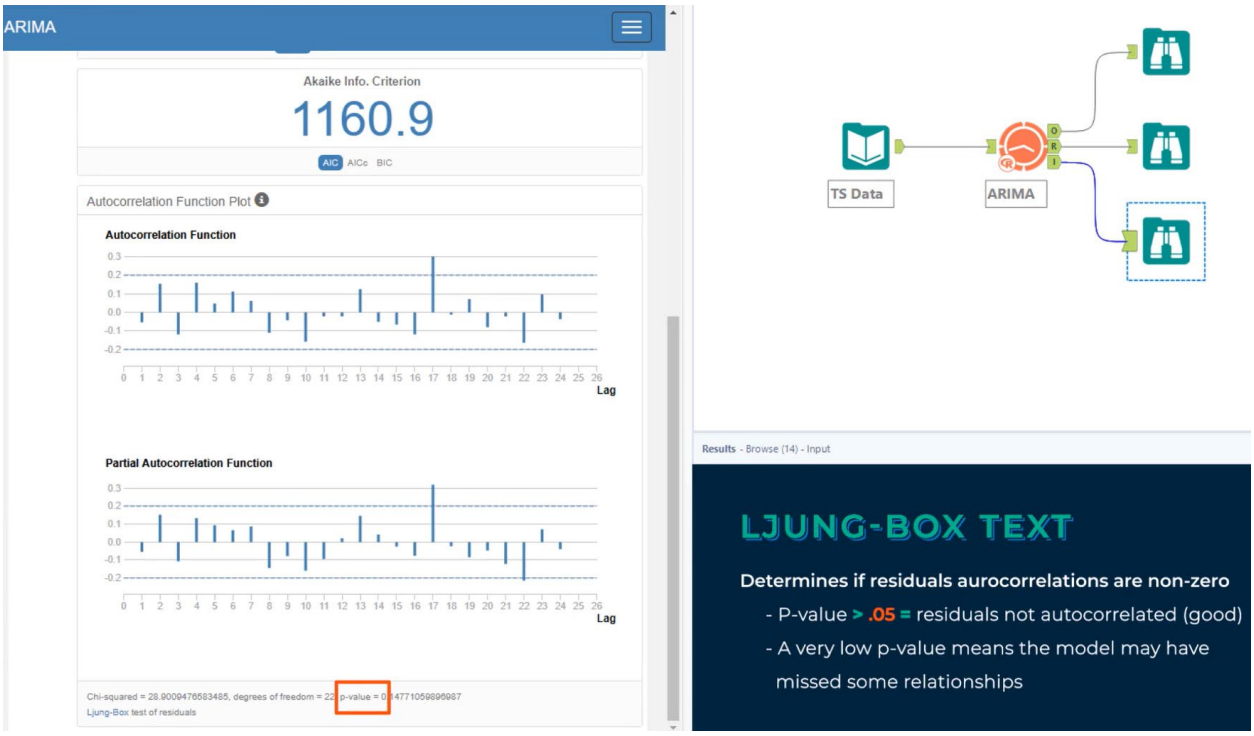
Box-Cox transformations stabilize VARIANCE

Differencing stabilizes the TREND

From ARIMA training


AR I MA
p d q
ALL PARAMETERS
0 = not present
d PARAMETER
1 = differenced once
2 = differenced twice
p, q PARAMETERS
1 = one coefficient
2 = two coefficeints
Summary of ARIMA Model
Method: ARIMA(0,1,1)(1,1,0)[12]
Number of periods in the seasonal cycle
Call:
auto.arima(Bookings)
seasonal components (1, 1, 0). These use P D Q (i.e, capitalized values)
Coefficients:
ma1 sar1
Value -0.671117 -0.449177
Std Err 0.077917 0.099353
sigma^2 estimated as 1578141.83851: log likelihood = -916.01114
Information Criteria:
AIC AICc BIC
1838.0223 1838.2553 1846.0408
In-sample error measures:
ME RMSE MAE MPE MAPE MASE ACF1
-14.6510224 1175.1059989 797.4171602 -3.2161678 16.0128095 0.5634056 -0.0071189
Ljung-Box test of the model residuals:
Chi-squared = 55.5596, df = 22, p-value = 9.9e-05
ARIMA (11) - Configuration
ETSworkflow.yxmd*
Required parameters
Model customization (optional)
Other options
Graphics Options
Model name
ARIMAExample
Select the target field
Bookings
Use covariates in model estimation? (Optional)
Select covariate field(s)
All
None
Year
Month
Bookings
Base forecast values on
Mean of covariates
Estimated change
Target field frequency
Hourly
Daily (all days)
Daily (weekdays only)
Weekly
Monthly
Quarterly
Annually
Other
COVARIATES
Predictor Variables
- One-hot encoded for categorical data
- e.g. A known holiday or annual sale
- Must be projected into the future
Leading Indicators
- Do not need to be forecast
- Precede a change in target variable
- e.g. A decline in interest rate today leading to more loan applications next month
Must be correlated with “noise”


ARIMA (11) - Configuration
Required parameters
Model customization (optional)
Other options
Graphics Options
Customize the parameters used for automatic model creation
The non-seasonal components
Alter the degree of first differencing...
The maximum order of the autoregressive component (p)
2
The maximum order of the moving average component (q)
2
The seasonal components
Alter the degree of seasonal differencing...
The maximum order of the seasonal autogressive component (P)
1
The maximum order of the seasonal movi
1
Information criteria for model selection
AIC
AICc
BIC
Do full enumeration of models (slow) instead of stepwise selection (faster)...
Allow drift
Use a Box-Cox transformation...
Completely user specified model...
Selected by default, this checkbox allows the final regression to have a constant value (similar to a y-intercept).



The “I” in ARIMA stands for “Integrated” which refers to the level of differencing.

Question 1

Differencing is associated with which of these models? Select all that apply.

- [x] ARIMA
- [x] ARIMA with Covariates
- [ ] ETS
- [ ] None of these

Question 2

What does a capital P of 0 (from P,D,Q) indicate in an ARIMA model?

- (●) No seasonal AR term
- ( ) None of these
- ( ) No seasonal MA term
- ( ) No AR term
- ( ) No MA term

Question 3

Which value is the most appropriate for this dataset's trend component?

- [x] None
- [ ] Additive
- [ ] Multiplicative

Question 4

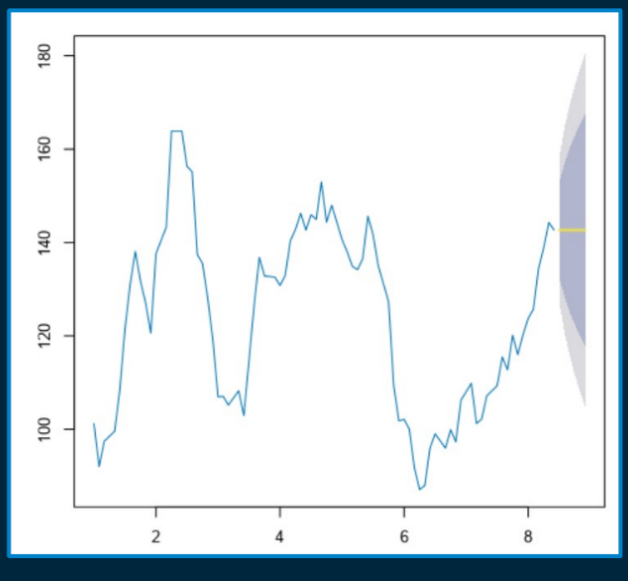
What is the purpose of a Box-Cox transformation?

- To make the data stationary
- To stabilize the variance
- To stabilize the mean
- None of these

Question 5

Which type of simple forecast is this?

- Dampened Drift
- Drift
- Seasonal Naïve
- Naïve

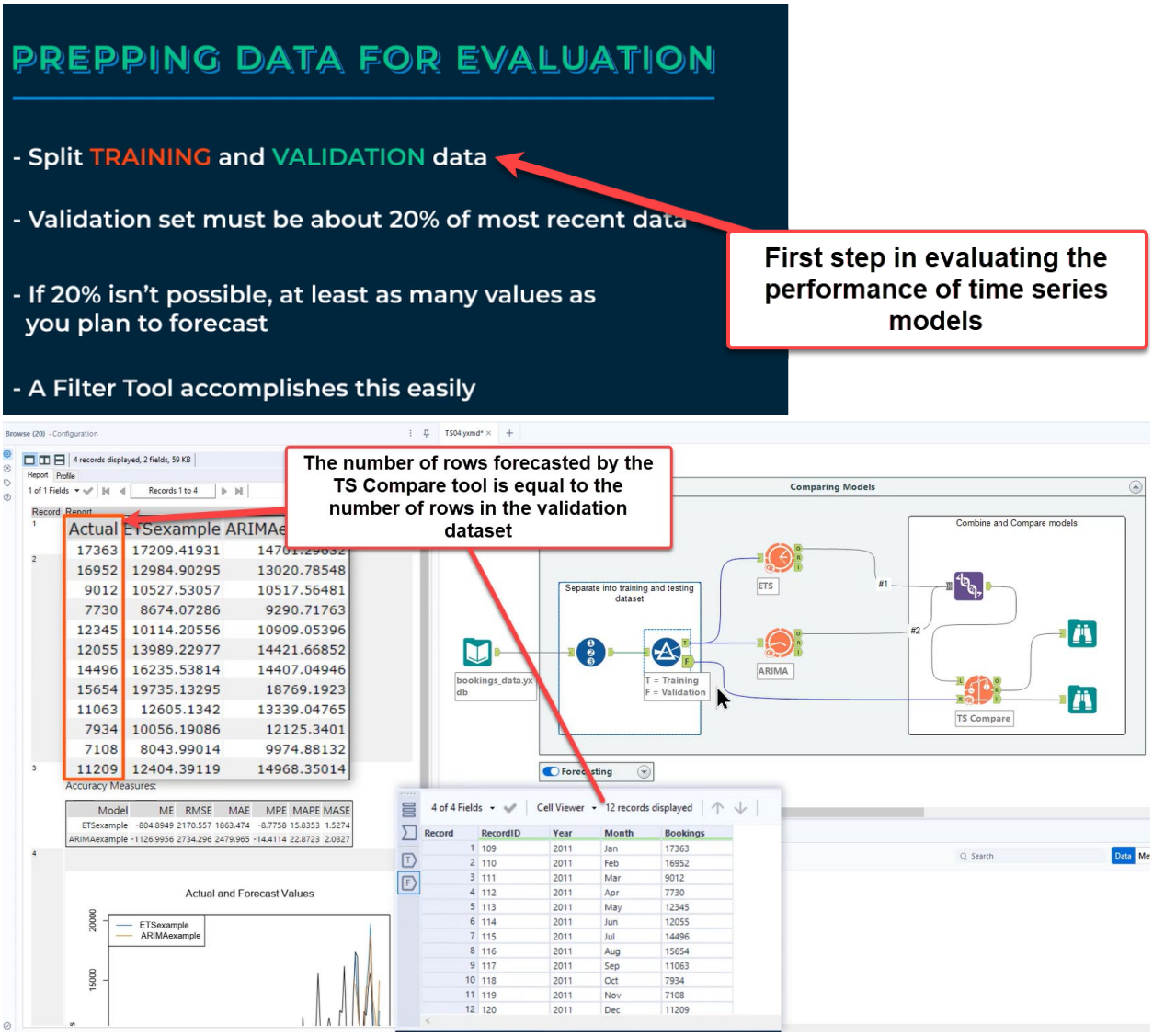


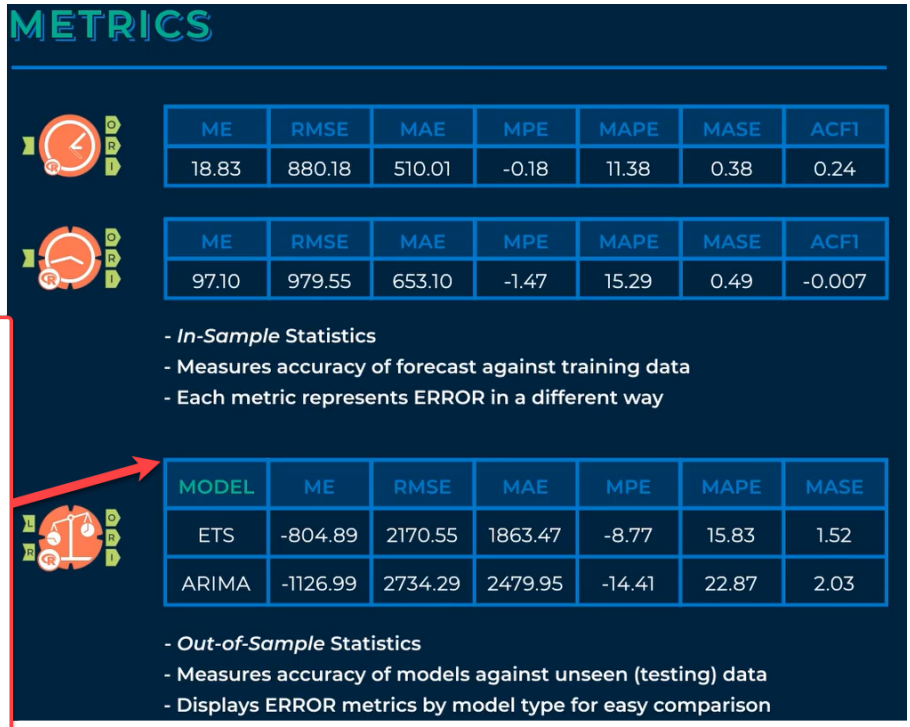
Question 6

What does an alpha close to 1 indicate in an ETS model?

- All datapoints are contributing equally to the calculation of level
- Only the most recent values are contributing to the calculation of seasonality
- There is no seasonal component in the data
- Only the most recent values are contributing to the calculation of level

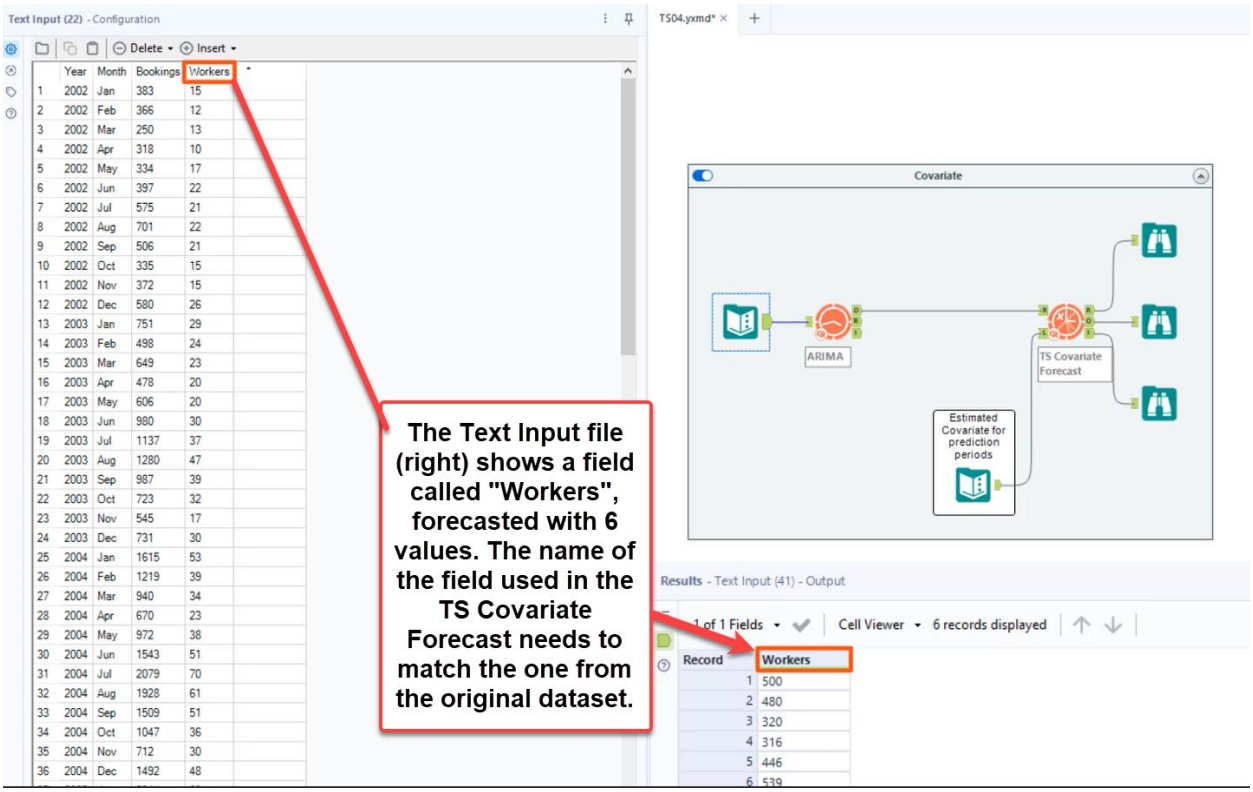
Selecting and Scaling Models



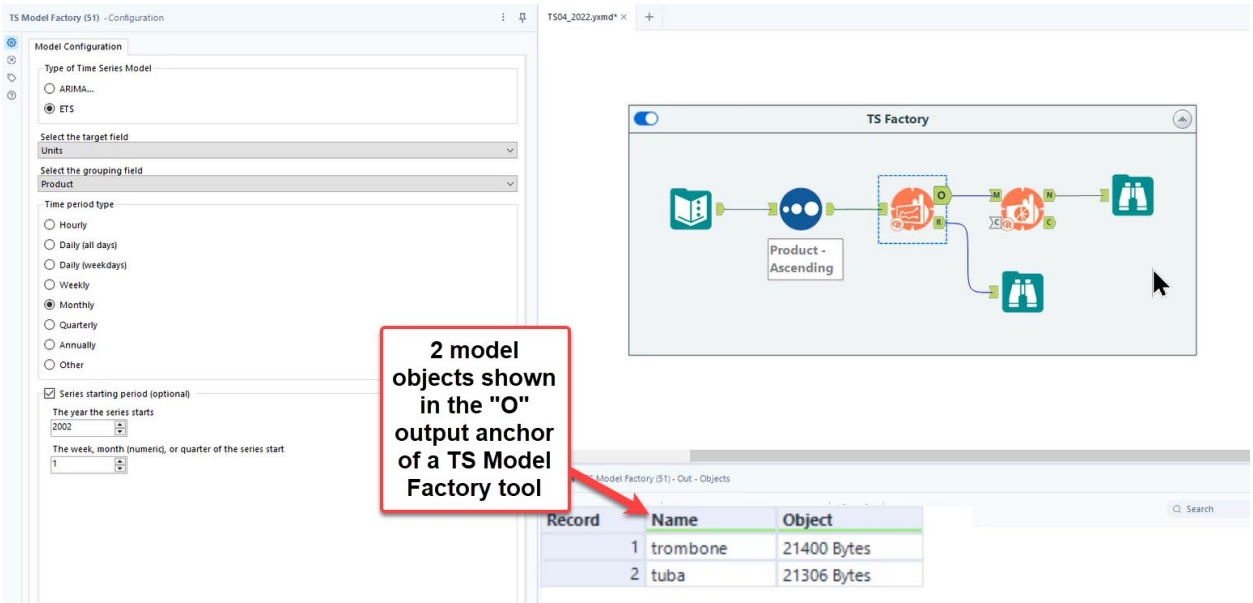


RMSE and MAE are Scale Dependent, and should not be compared between models that have datasets with different units. MPE and MAPE, which are percent errors, can be compared across units. Scaled error terms, such as MASE, are less affected by outliers. For any of the metrics, the lower the absolute value, the better the performance.

Remember that the initial model used to create an ETS or ARIMA contains only the training data. Once a model is selected, create a new, full model that has ALL of the data (i.e., both training and validation) before using a TS Forecast tool (unless you made an ARIMA model that uses covariate(s)).



The TS Model Factory tool is used to create more than one time series model. An example is grouping by a field named Products, where there is more than one product in the field. The data, as usual, need to be sorted ascending. Note that it is ideal, but not necessary, to have the same number of records for each group.



The TS Forecast Factory tool is similar to the TS Forecast tool, except that the former has an optional input anchor, to be used when forecasting with covariates.

Question 1

How is MAPE different from RMSE?

- [x] MAPE can be used to compare models created from datasets with different units while RMSE cannot.
- [ ] None of these.
- [ ] RMSE can be used to compare models created from datasets with different units while MAPE cannot.

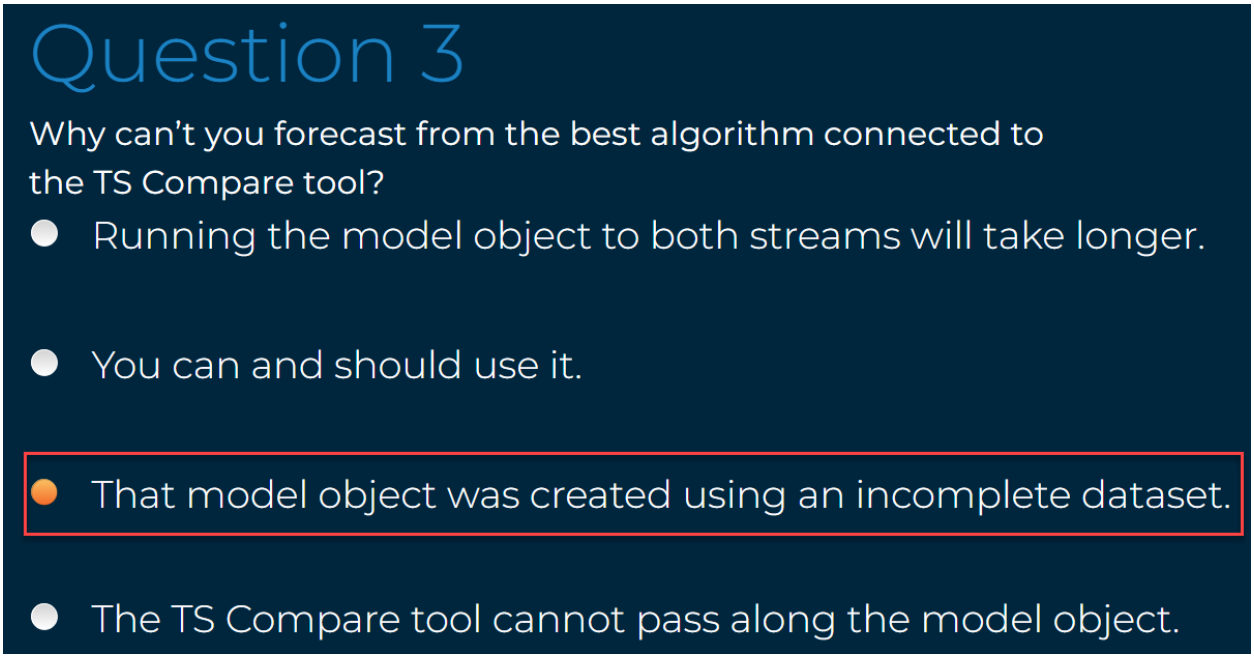
Question 2

What is the difference between the RMSE from the algorithm, and the TS Compare tool?

ME	RMSE	MAE	MPE	MAPE	MASE	AFC1
-194.24	993.55	539.26	-4.601	11.83	0.410	0.091

MODEL	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-804.89	2170.55	1863.47	-8.77	15.83	1.52
ARIMA	-1126.99	2734.29	2479.95	-14.41	22.87	2.03

- [ ] There is no difference.
- [ ] One is a prediction and the other is a forecast.
- [ ] One is calculated using an estimated standard deviation and the other uses the computed standard deviation for the dataset.
- [x] One is an in-sample statistic and the other measures performance on unseen data.



Note that categorical variable time series can be used as covariate series in your model and forecast. They will be one-hot encoded by Designer so that they take a numerical form in the analysis.

Optimization

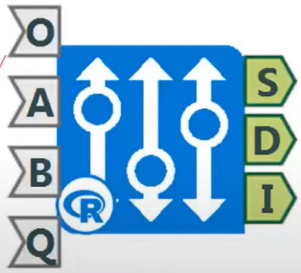
Alteryx Live Training: Prescriptive Optimization

INPUT ANCHORS FOR MATRIX INPUT MODE

Objective

- Required Fields
  - Variable – can be logical
  - Coefficient (prices of variables)
  - Lower Boundary (named 'lb')
  - Upper Boundary (named 'ub')
  - Type – Three values to populate
    - C – Continuous
    - I – Integer
    - B – Binary

Record #	variable	coefficient	type	lb	ub
1	Pants	50	C	200	Inf
2	Jackets	40	C	0	400



Alteryx Live Training: Prescriptive Optimization

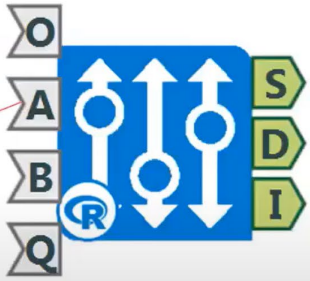
INPUT ANCHORS FOR MATRIX INPUT MODE

Left-hand-side constraints (LHS)

- x+1.5y<=750
- 2x+y<=1000

- Left-side of equalities go into this input
- Keep variables as rows and constraints as fields
- Variable field must be named 'variable'

Record #	variable	CottonSqFt	PolyesterSqFt
1	Pants	1	2
2	Jackets	1.5	1



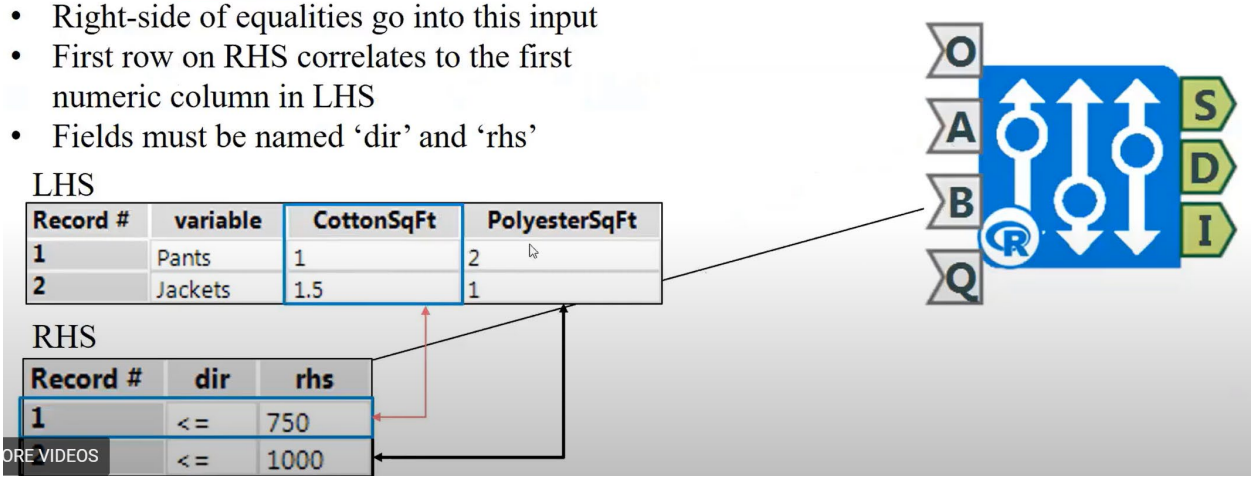
Alteryx Live Training: Prescriptive Optimization

INPUT ANCHORS FOR MATRIX INPUT MODE

Right-hand-side constraints (RHS)

- x+1.5y<=750
- 2x+y<=1000

- Right-side of equalities go into this input
- First row on RHS correlates to the first numeric column in LHS
- Fields must be named 'dir' and 'rhs'



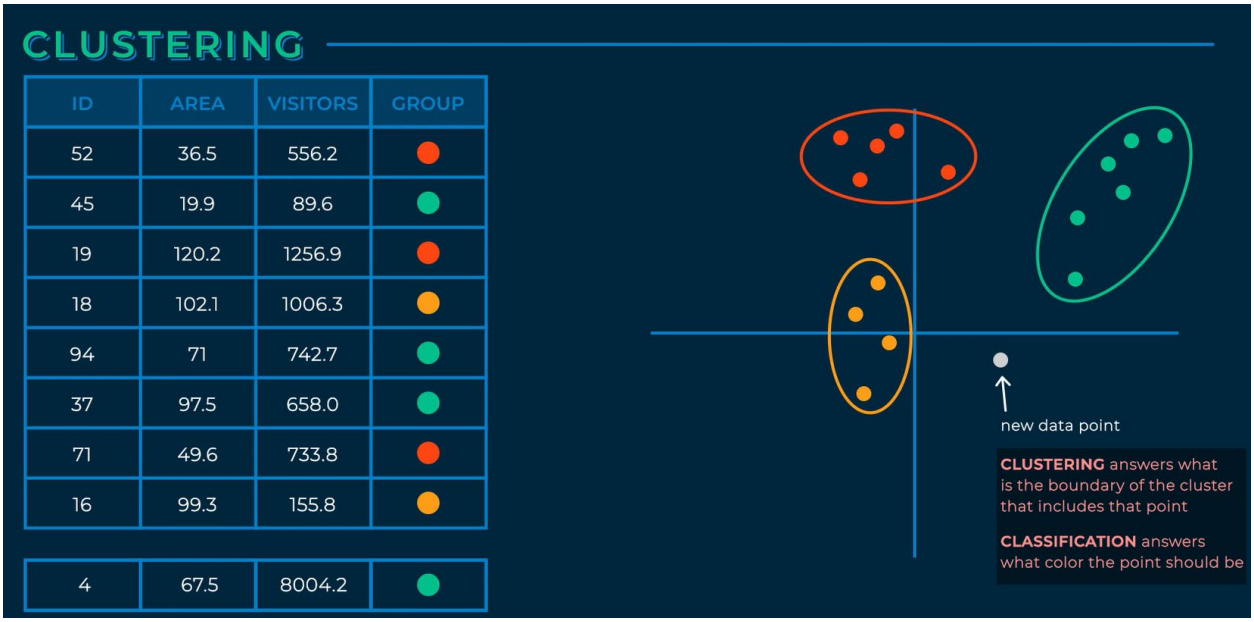
LHS

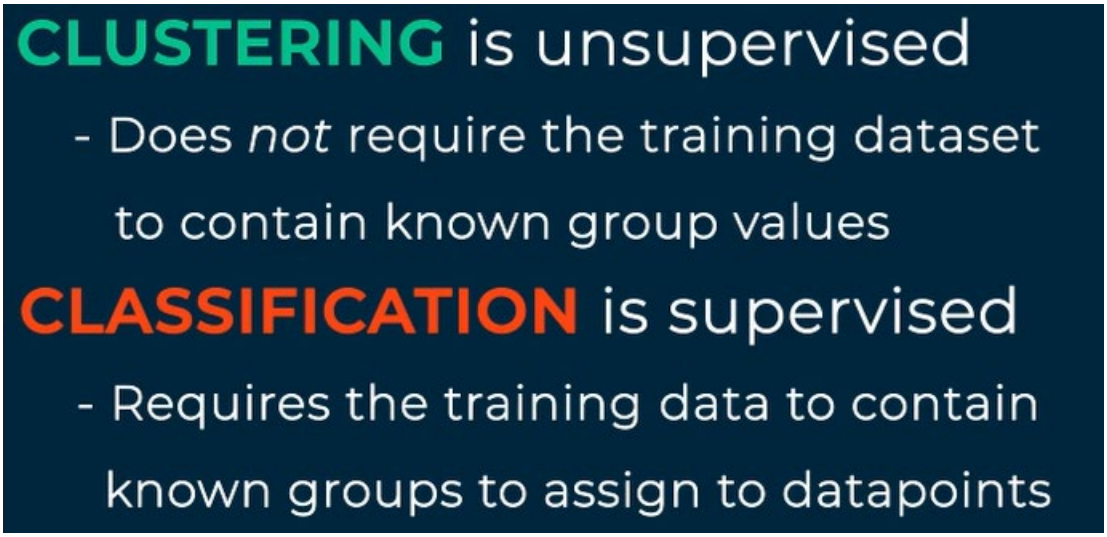
Record #	variable	CottonSqFt	PolyesterSqFt
1	Pants	1	2
2	Jackets	1.5	1

RHS

Record #	dir	rhs
1	<=	750
	<=	1000

Predictive Grouping

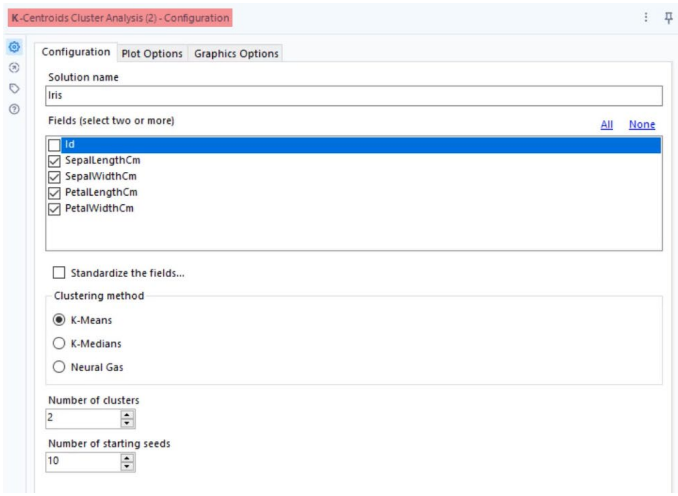




Principal component analysis (PCA) is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

CLUSTERS

Groupings of similar datapoints and separation of dissimilar datapoints.



Clustering Workflow.yxmd*

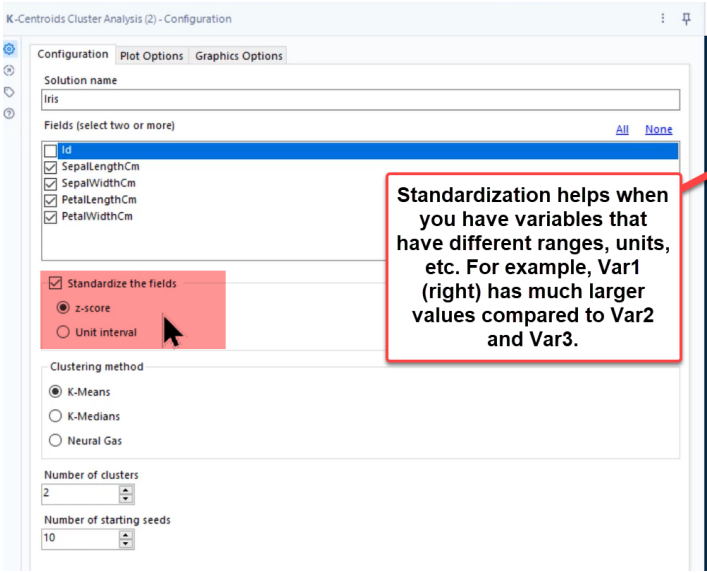
Select Variables

Only numeric columns are displayed

Minimum of 2 variables selected

Cannot contain Null values

Sensitive to outliers



Clustering Workflow.yxmd*

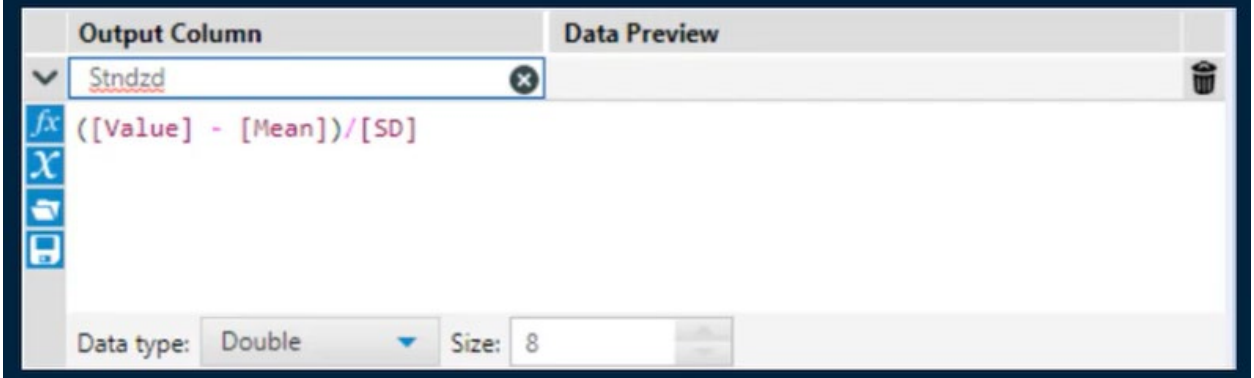
Standardization

Standardization helps when you have variables that have different ranges, units, etc. For example, Var1 (right) has much larger values compared to Var2 and Var3.

Equalize the units and/or range of values

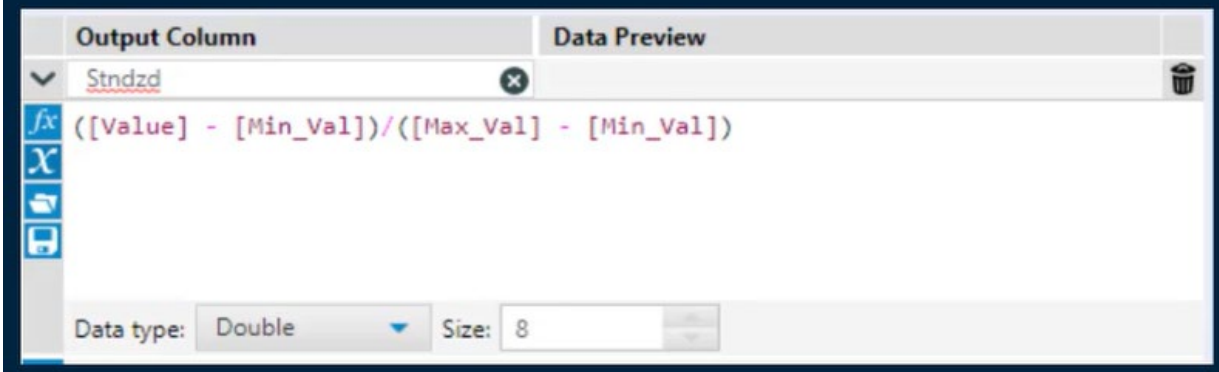
Var1	Var2	Var3
1256	12.6	8.6
1348	13.8	12
1096	14.4	13.1
973	8.9	13.6
1120	10.2	10.3
854	11.5	7.5

Z-Score

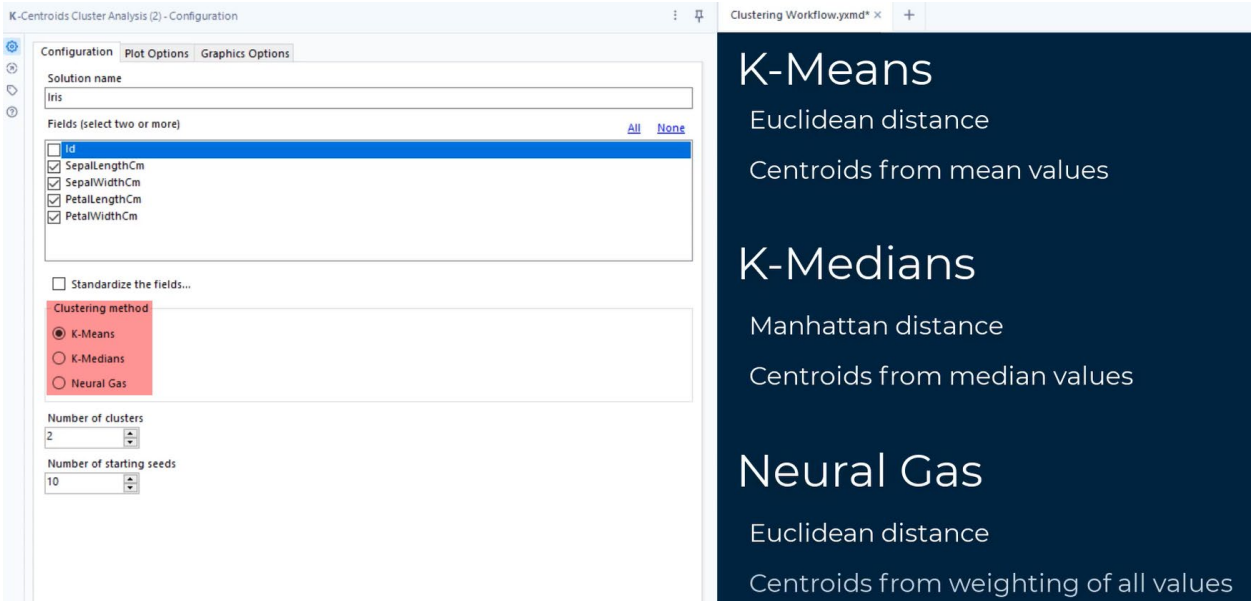


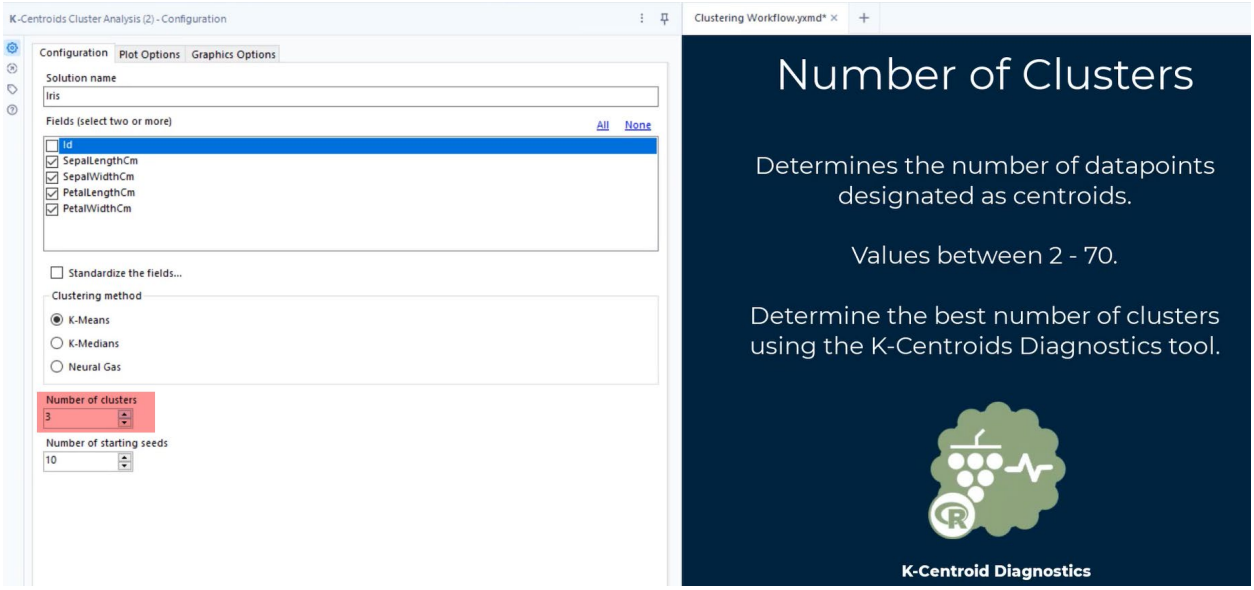
Results in a variable with a Mean of 0 & SD of 1

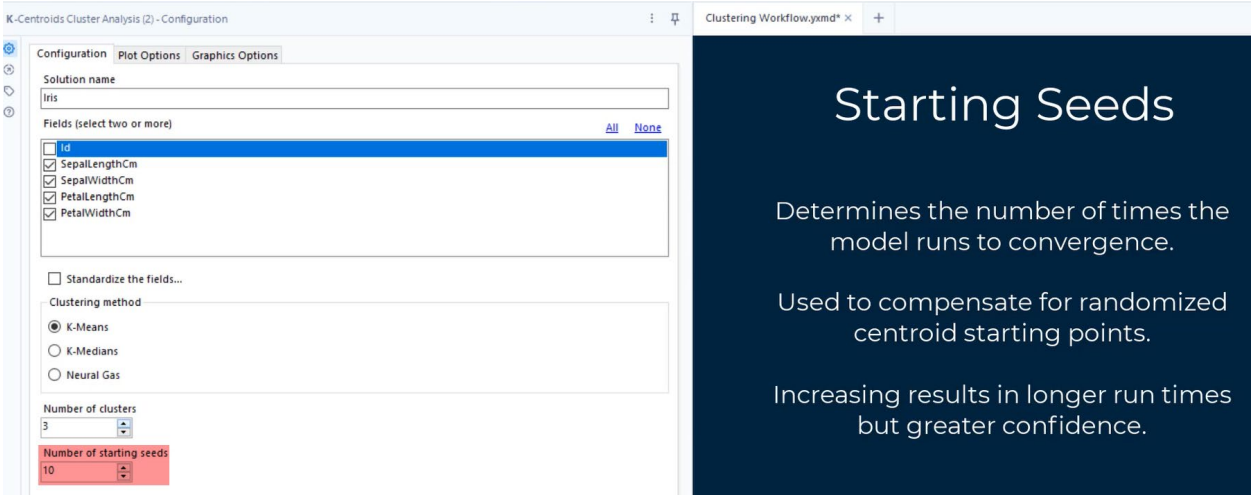
Unit Interval

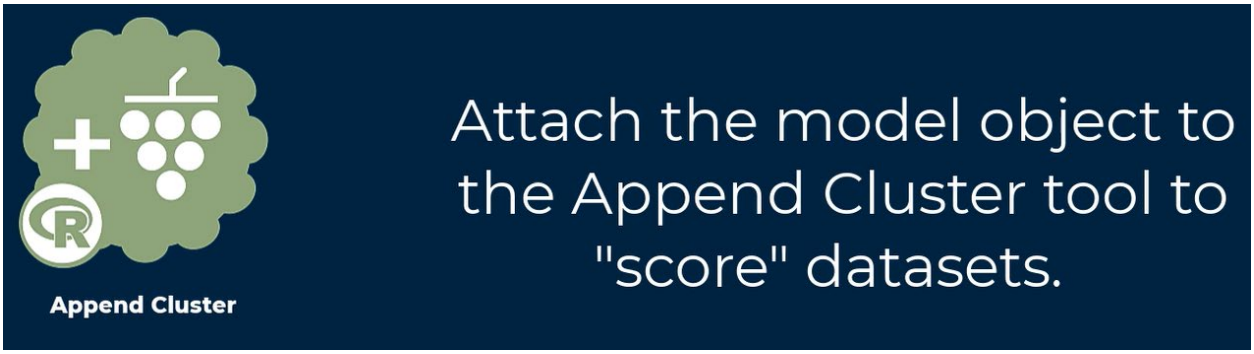


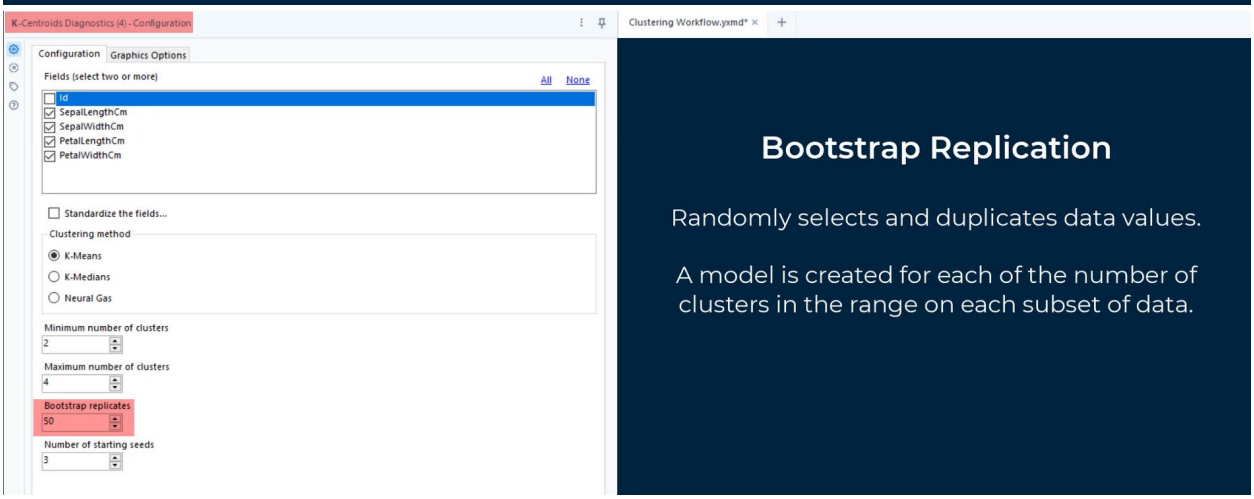
Results in a variable with a range of 0 - 1

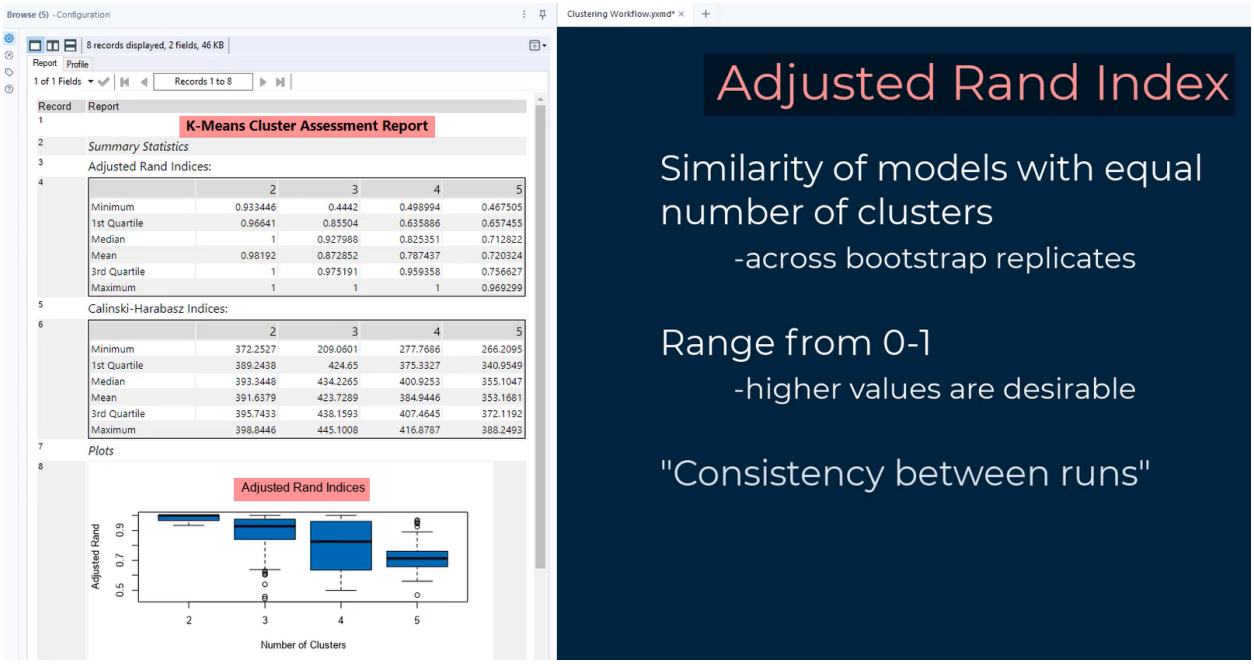

K-Centroids Cluster Analysis (2) - Configuration
Clustering Workflow.yxmd*
Configuration
Plot Options
Graphics Options
Solution name
Iris
Fields (select two or more)
All
None
Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Standardize the fields...
Clustering method
K-Means
K-Medians
Neural Gas
Number of clusters
2
Number of starting seeds
10
K-Means
Euclidean distance
Centroids from mean values
K-Medians
Manhattan distance
Centroids from median values
Neural Gas
Euclidean distance
Centroids from weighting of all values


K-Centroids Cluster Analysis (2) - Configuration
Clustering Workflow.yxmd*
Configuration
Plot Options
Graphics Options
Solution name
Iris
Fields (select two or more)
All
None
Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Standardize the fields...
Clustering method
K-Means
K-Medians
Neural Gas
Number of clusters
3
Number of starting seeds
10
Number of Clusters
Determines the number of datapoints designated as centroids.
Values between 2 - 70.
Determine the best number of clusters using the K-Centroids Diagnostics tool.
K-Centroid Diagnostics


K-Centroids Cluster Analysis (2) - Configuration
Clustering Workflow.yxmd*
Configuration
Plot Options
Graphics Options
Solution name
Iris
Fields (select two or more)
All
None
Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Standardize the fields...
Clustering method
K-Means
K-Medians
Neural Gas
Number of clusters
3
Number of starting seeds
10
Starting Seeds
Determines the number of times the model runs to convergence.
Used to compensate for randomized centroid starting points.
Increasing results in longer run times but greater confidence.







	2	3	4	5
Minimum	0.933446	0.4442	0.498994	0.467505
1st Quartile	0.96641	0.85504	0.635886	0.657455
Median	1	0.927988	0.825351	0.712822
Mean	0.98192	0.872852	0.787437	0.720324
3rd Quartile	1	0.975191	0.959358	0.756627
Maximum	1	1	1	0.969299

	2	3	4	5
Minimum	372.2527	209.0601	277.7686	266.2095
1st Quartile	389.2438	424.65	375.3327	340.9549
Median	393.3448	434.2265	400.9253	355.1047
Mean	391.6379	423.7289	384.9446	353.1681
3rd Quartile	395.7433	438.1593	407.4645	372.1192
Maximum	398.8446	445.1008	416.8787	388.2493

Calinski-Harabasz Index

Definition of the clusters

-similarity of like datapoints vs differences between clusters

Ratio:

-numerator = separation of clusters

-denominator = tightness in clusters

Higher values are desirable

Report | Profile

1 of 1 Fields | Records 1 to 8

3rd Quartile	1	0.975191	0.959358	0.756627
Maximum	1	1	1	0.969299

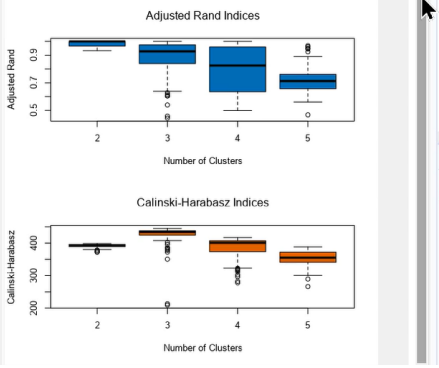
5 Calinski-Harabasz Indices:

6

	2	3	4	5
Minimum	372.2527	209.0601	277.7686	266.2095
1st Quartile	389.2438	424.65	375.3327	340.9549
Median	393.3448	434.2265	400.9253	355.1047
Mean	391.6379	423.7289	384.9446	353.1681
3rd Quartile	395.7433	438.1593	407.4645	372.1192
Maximum	398.8446	445.1008	416.8787	388.2493

7 *Plots*

8

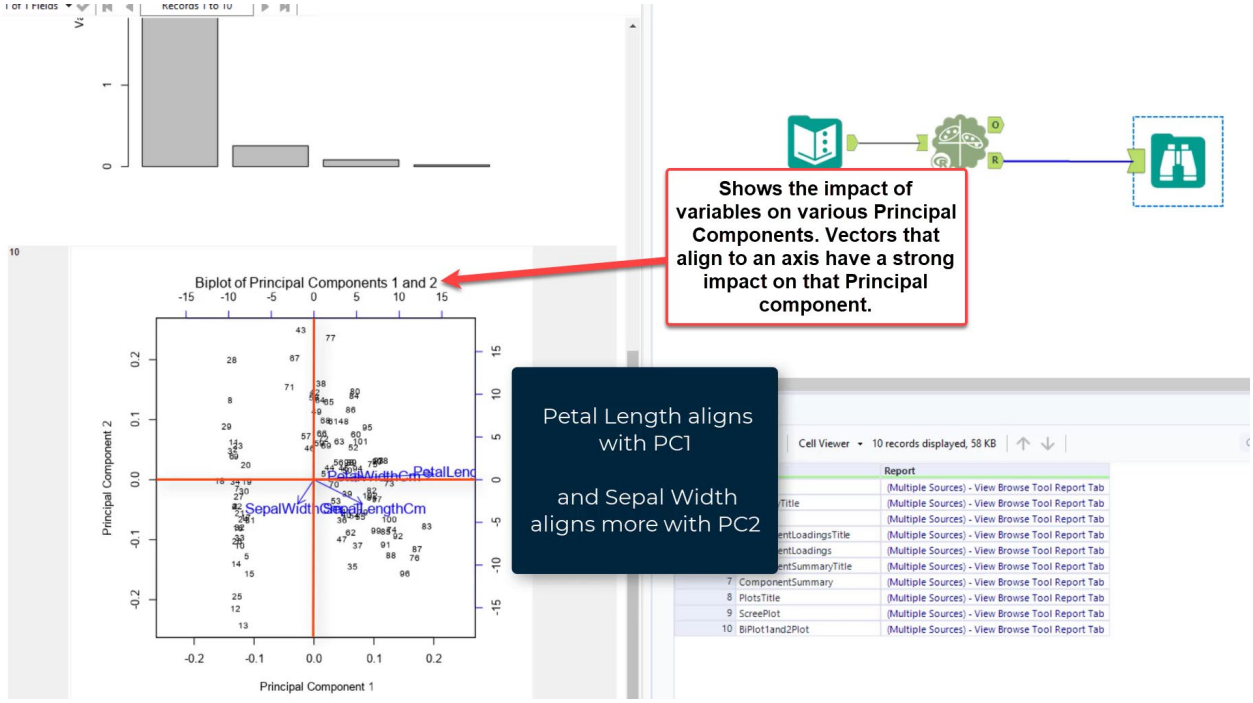


In this example, 2 or 3 clusters seem to be the best options.

Results - Browse (5) - Input

2 of 2 Fields | Cell Viewer | 8 records displayed, 46 KB | Search | Data | Metadata

Record	Desc	Report
1	Title	(Multiple Sources) - View Browse Tool Report Tab
2	SummaryTitle	(Multiple Sources) - View Browse Tool Report Tab
3	RandIndicesTitle	(Multiple Sources) - View Browse Tool Report Tab
4	RandIndicesTable	(Multiple Sources) - View Browse Tool Report Tab
5	CHTitle	(Multiple Sources) - View Browse Tool Report Tab
6	CHTable	(Multiple Sources) - View Browse Tool Report Tab
7	PlotsTitle	(Multiple Sources) - View Browse Tool Report Tab
8	Plots	(Multiple Sources) - View Browse Tool Report Tab



CONSIDERATIONS

Using PCA to create models requires the same process be applied to new datasets before predictions can be made.

PCA is a powerful tool but not a universal solution.

Be mindful of potential pitfalls.

- May interpret noisiest data as most important.
- PCA may use correlations not based in reality.

Proper data investigation is essential before performing PCA.

Question 1

Which of the following are potential issues when using PCA?
Select all that apply.

- [x] Noisy data is interpreted as important
- [x] Interpretability is lost
- [ ] Variables with a larger range of values are interpreted as important
- [ ] Variables containing little information are ignored

This is also true (assuming scaling/standardization hasn't taken place)

Question 2

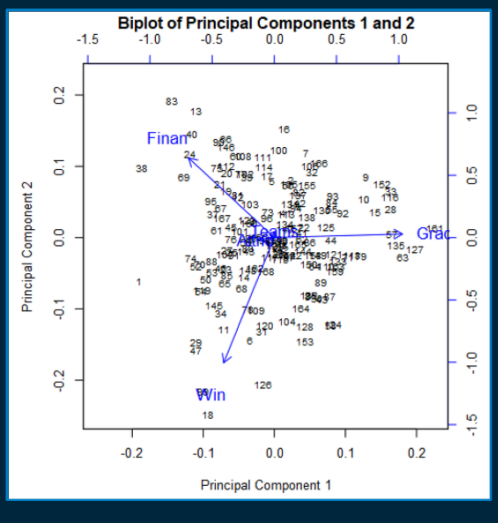
PCA performs dimensional reduction by:

- ○ Reducing Variance
- ○ N/A. It increases the number of columns
- ● Combining information from all selected columns
- ○ Removing columns

Question 3

Based on the Biplot pictured, which factor has the smallest impact on Principal Component 1?

- Finan
- Win
- Not enough info provided
- Grad



Covariance is always measured in two dimensions. If you are dealing with more than two variables, the most efficient way to make sure you get all possible covariance values is to put them into a matrix (hence, covariance matrix). In a covariance matrix, the diagonal is the variance for each variable, and the values across the diagonal are a mirror for one another because each combination of variables is included in the matrix twice. This is a square, symmetric matrix.

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 0.67 & 0.55 \\ 0.55 & 0.25 \end{bmatrix} \end{array}$$

In this example, the variance of variable A is 0.67, and the variance of the second variable is 0.25. The covariance between the two variables is 0.55, which is mirrored across the main diagonal of the matrix.