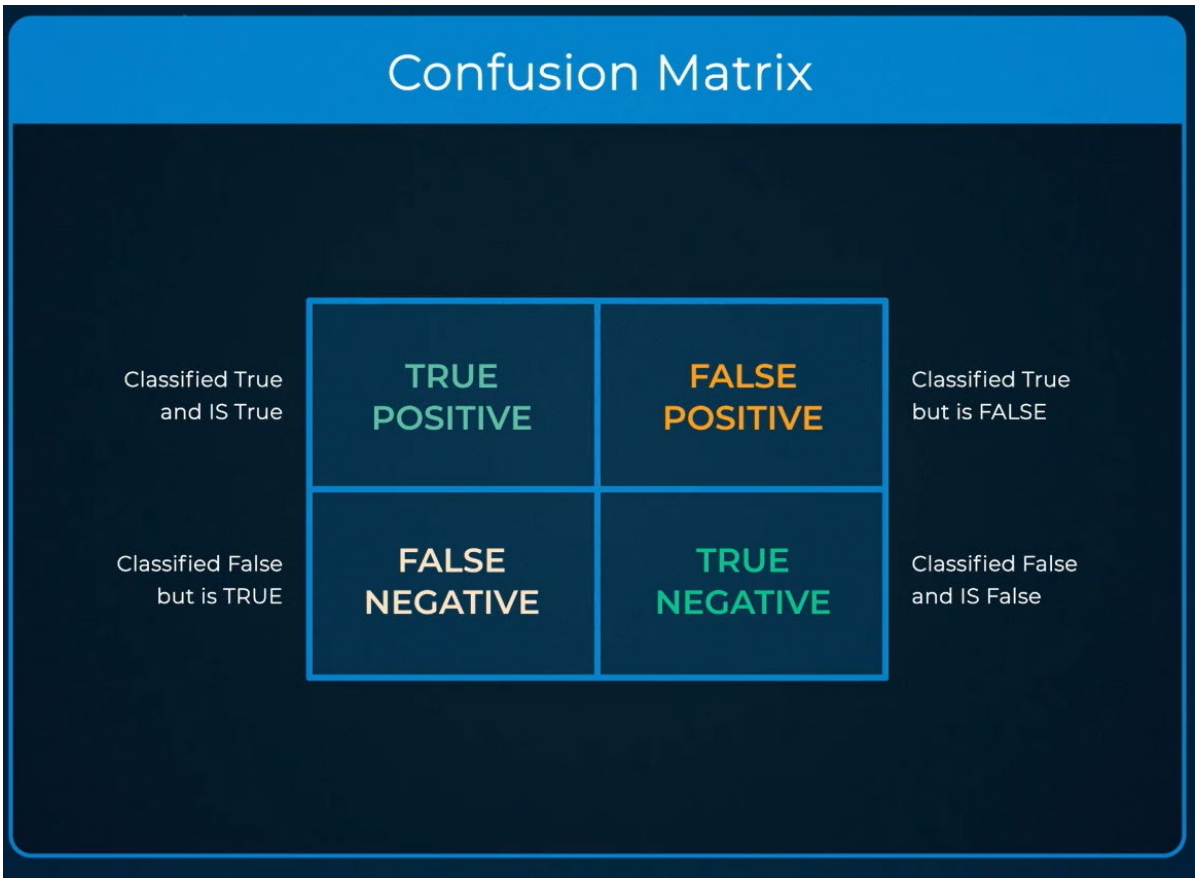
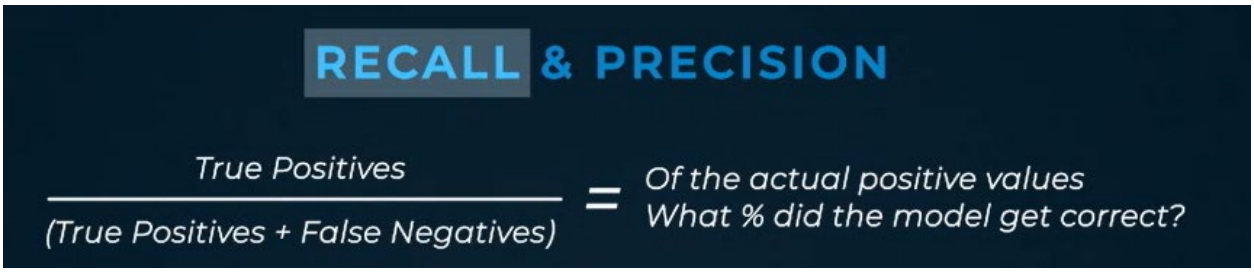
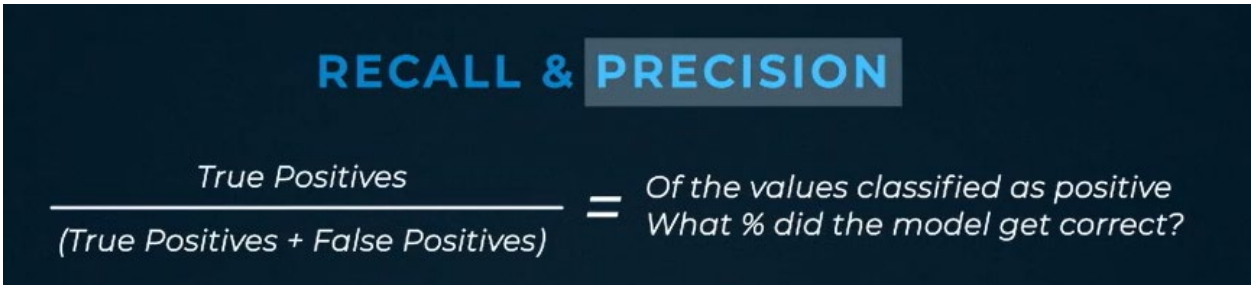
Expert Exam

Time Series, Regression, Data Investigation, Classification, and Optimization

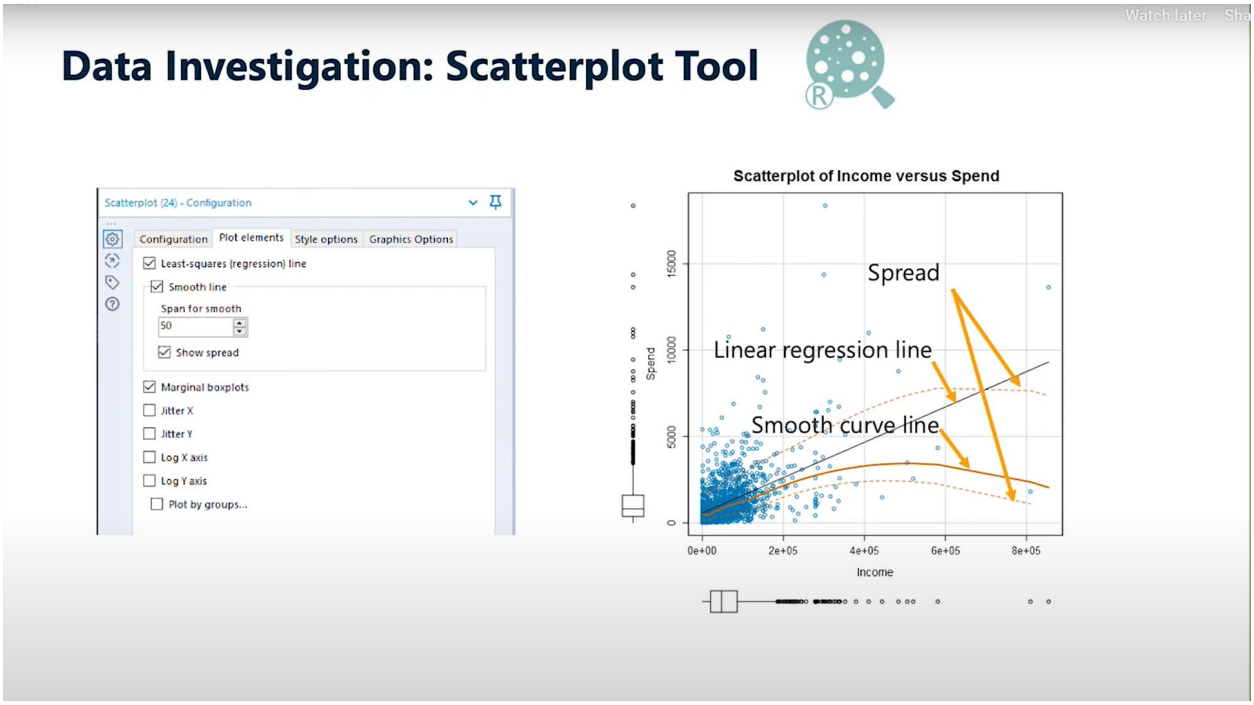
1. How to Interpret Pr(>|t|) in Regression Model Output in R
2. How to Interpret Significance Codes in R
3. How to Calculate a P-Value from a T-Test By Hand
4. Live Training: Regression Modeling
5. Mastering Time Series Analysis & Forecasting in Alteryx
6. Confusion Matrix







7. Holdouts and Cross Validation: Why the Data Used to Evaluate your Model Matters
8. Coefficient of determination
9. How To: Complete Data Preparation And Investigation For Predictive Modeling
10. Cross Validation
11. Binomial coefficient
12. Knapsack problem
13. Ljung-Box Test: Definition + Example
14. Pearson Correlation vs. Hoeffding's D statistic
15. Scatterplot



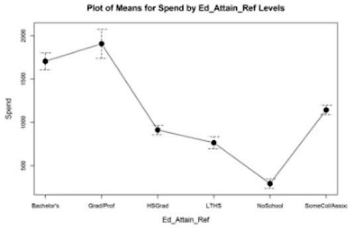
16. Plot of Means

Data Investigation: Plot of Means Tool

Compare two variables:

- Response field: numeric or binary categorical vs. categorical field
- For example:
  - Spend versus educational attainment

One Output



Graph

17. Action

Action

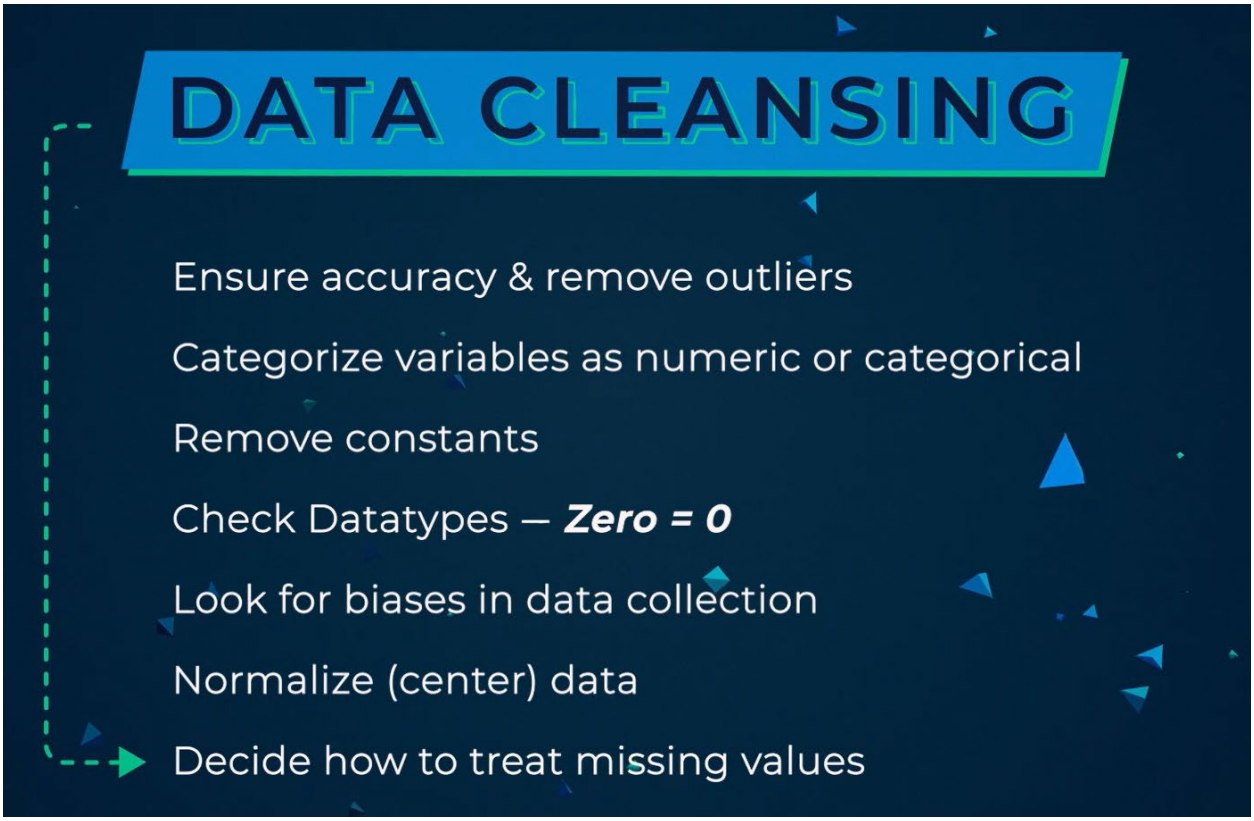
- Updates workflow with values from the interface
- Action tool has many different configurations
  - Configurations available vary by tool
- Download actions available on the help menu

Accepts output connections from interface tools with the same icon

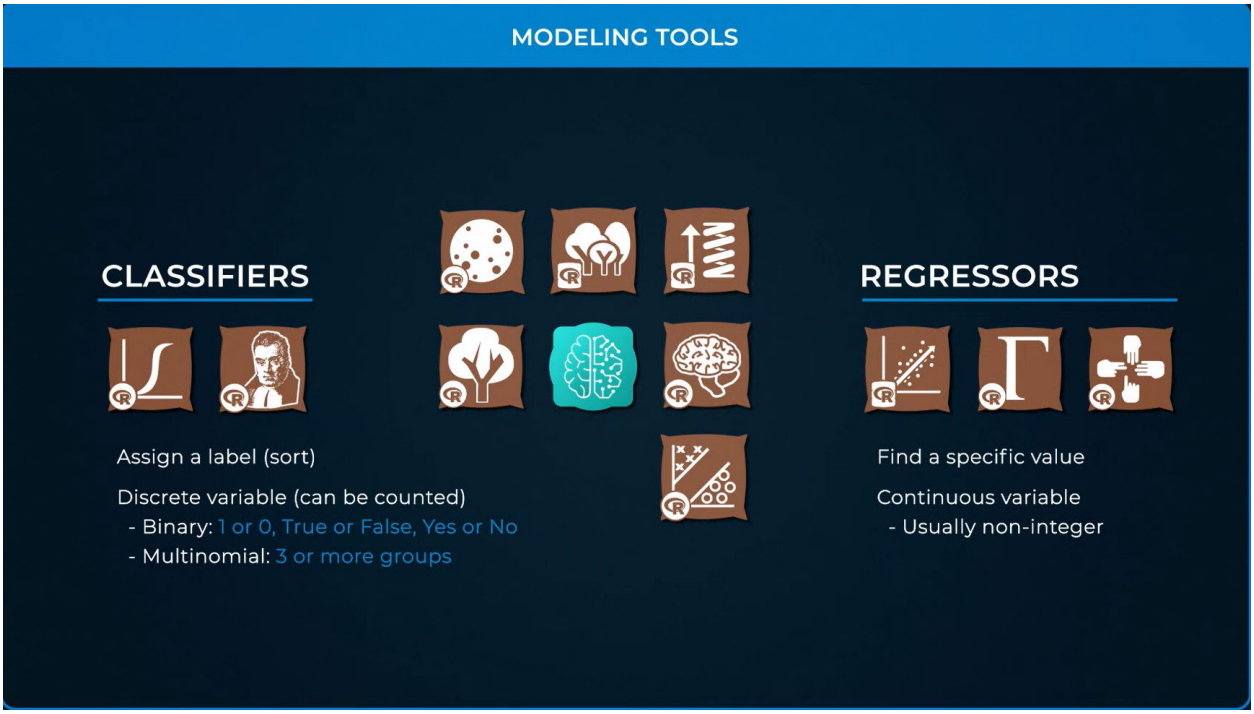
Only accepts incoming connections from the conditional tool

Connect lightning bolt to a workflow tool

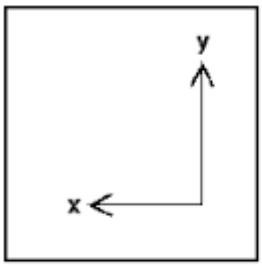
18. A model is a mapping of the relationships in your data.
19. Third step of creating a model is *Data Cleansing*.



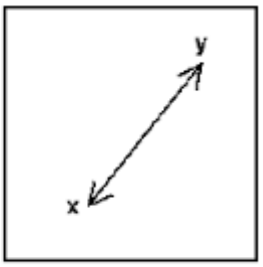
20. Modeling tools. The models in the middle can provide predictions for both categorical and continuous variables.



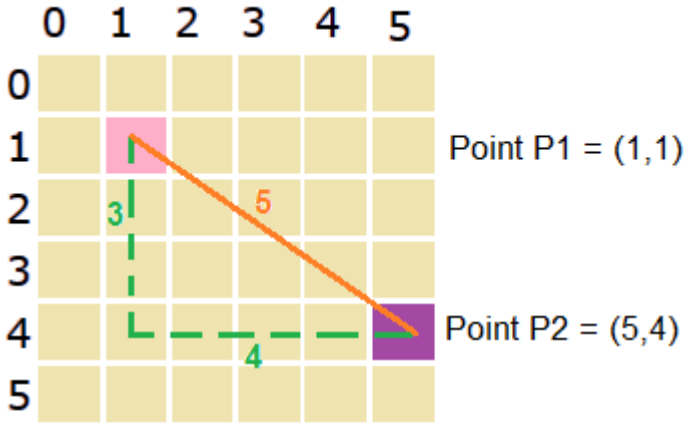
21. Manhattan vs Euclidean (aka “as the crow flies”):



Manhattan



Euclidean



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

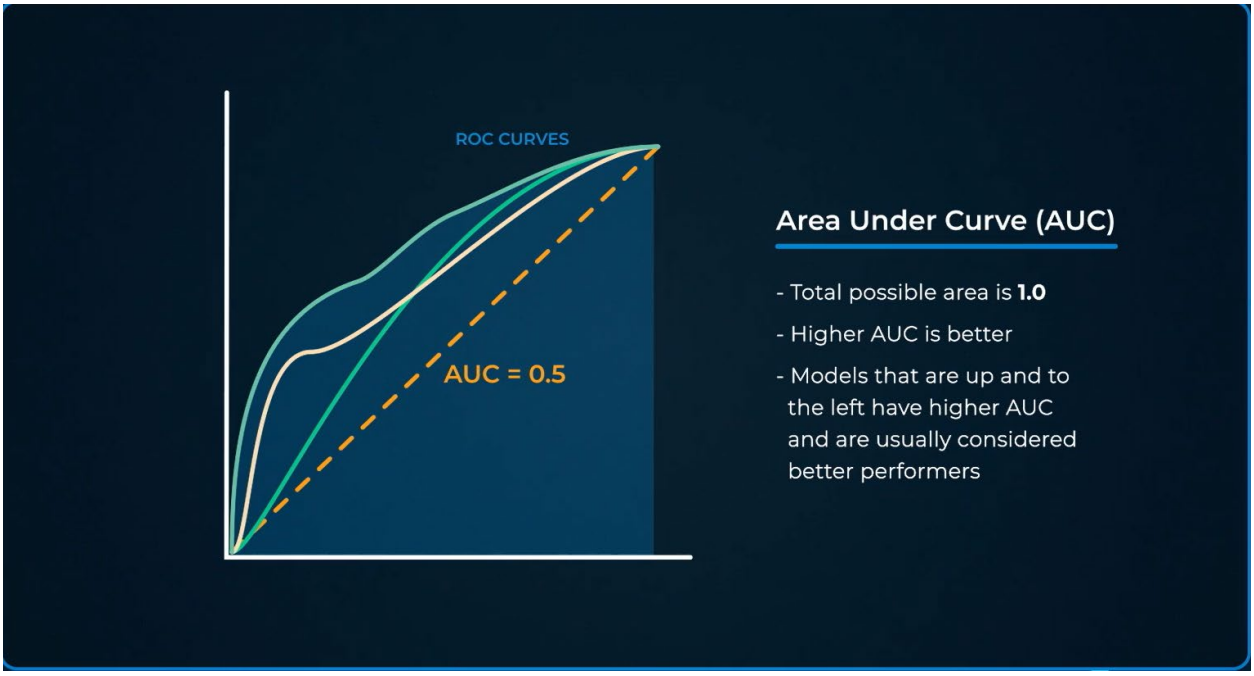
22. Collinearity

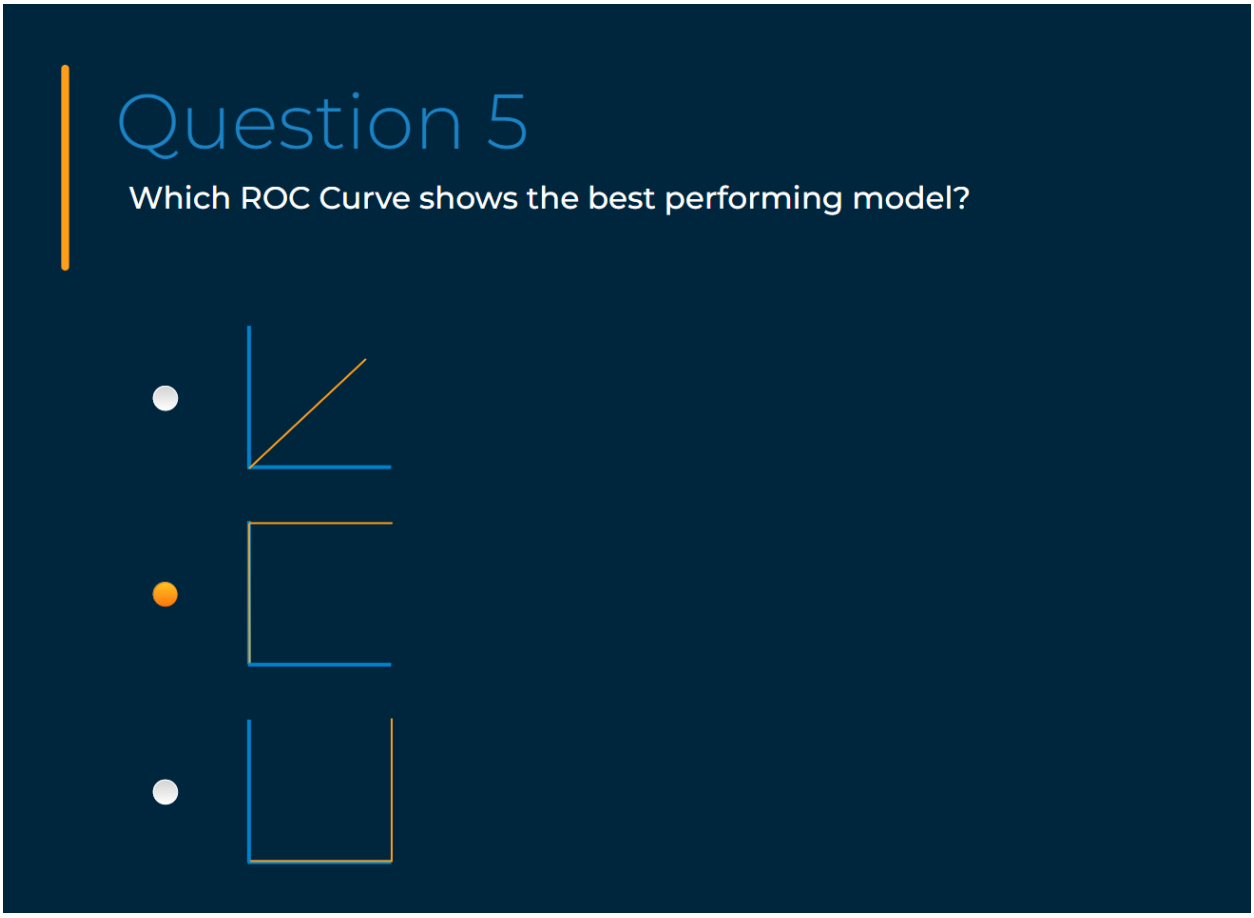
Question 2

What is collinearity?

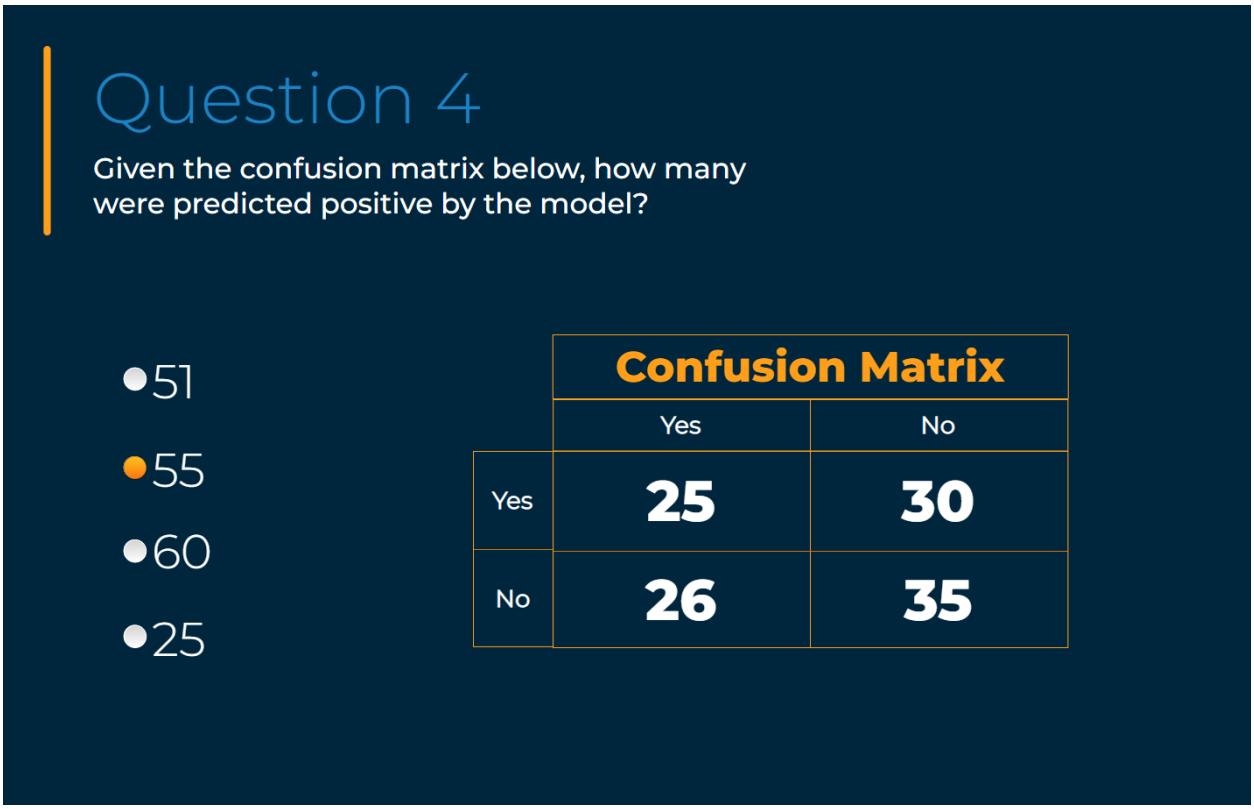
- When more than one model fits the data equally well.
- When multiple independent variables describe the same information.
- The relationship between two variables.
- None of these.

23. Area Under Curve (AUC)





24. Confusion Matrix question (predicted in rows; actual in columns).



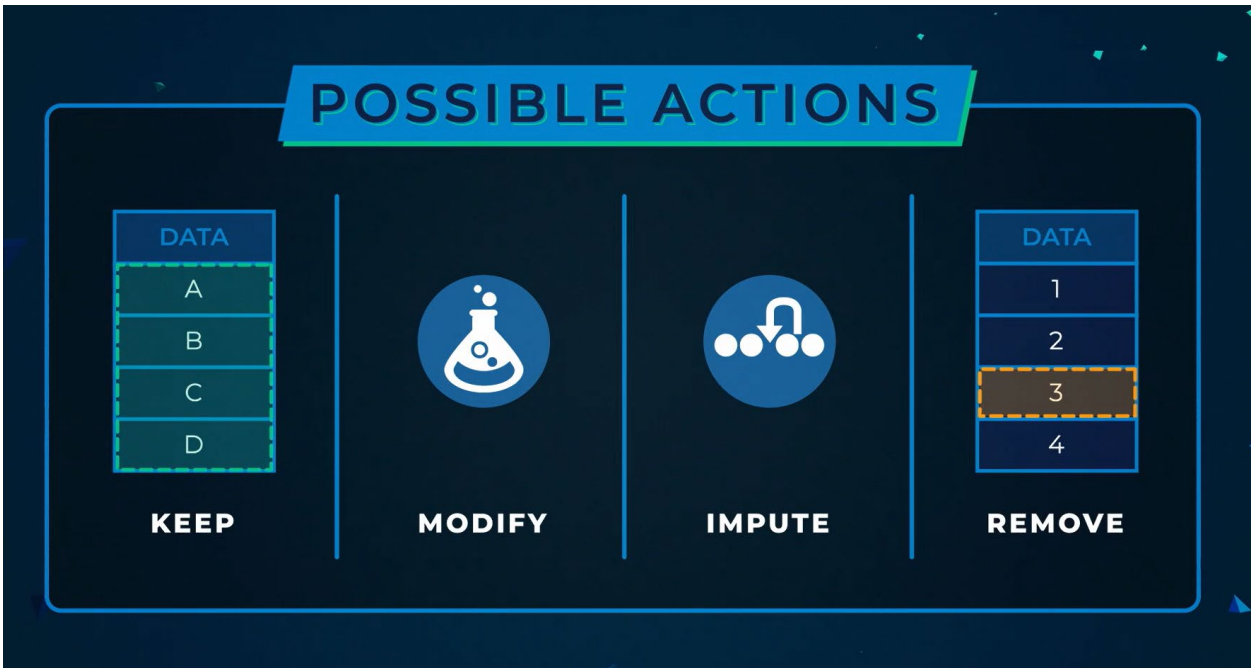
25. Predictive Analytics question

Question 6

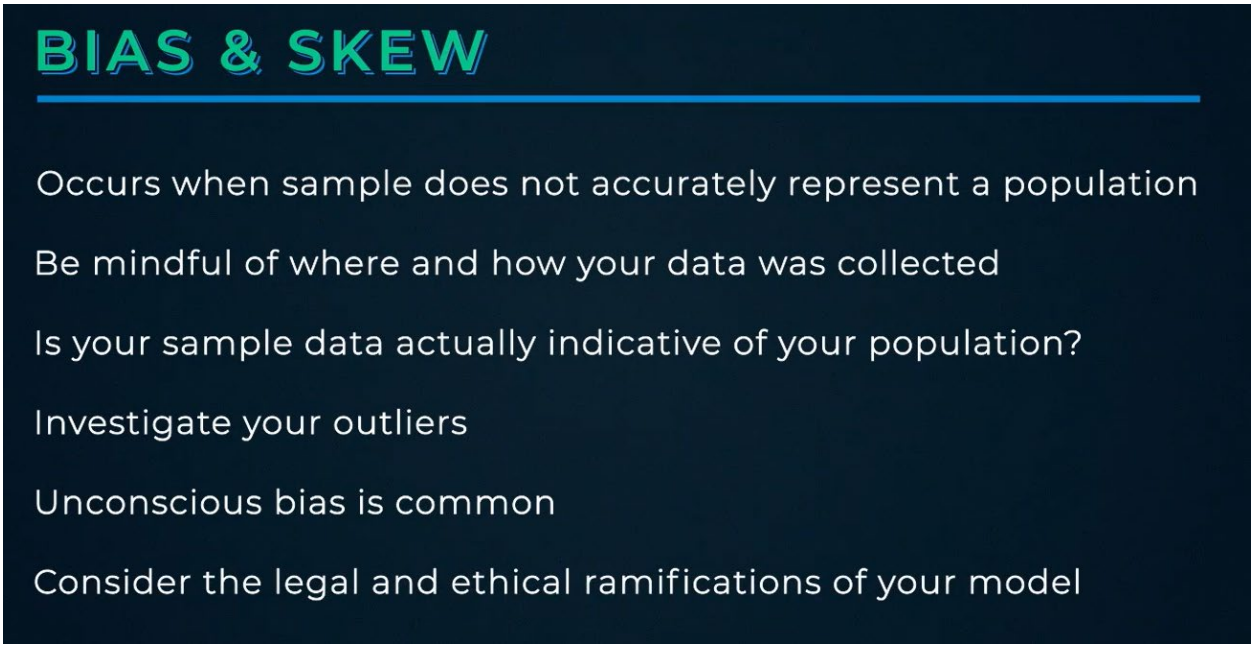
Which of these is the function for accuracy?

- TP+FP/TP+FP+FN+TN
- TP/TP+FN
- TP/TP+FP
- None of these

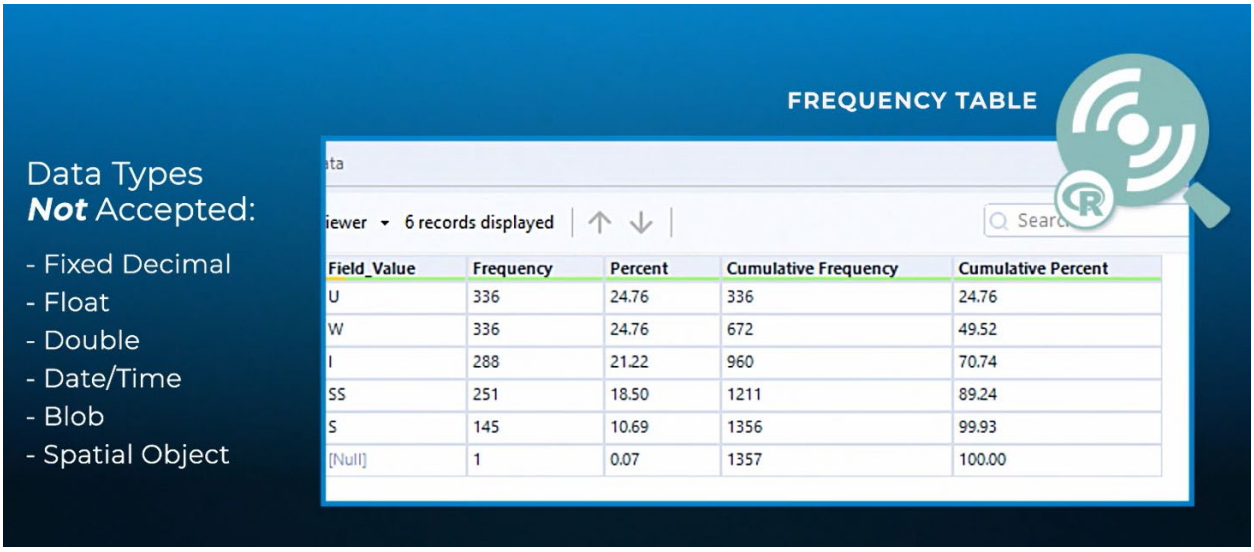
26. Data Investigation courses of action



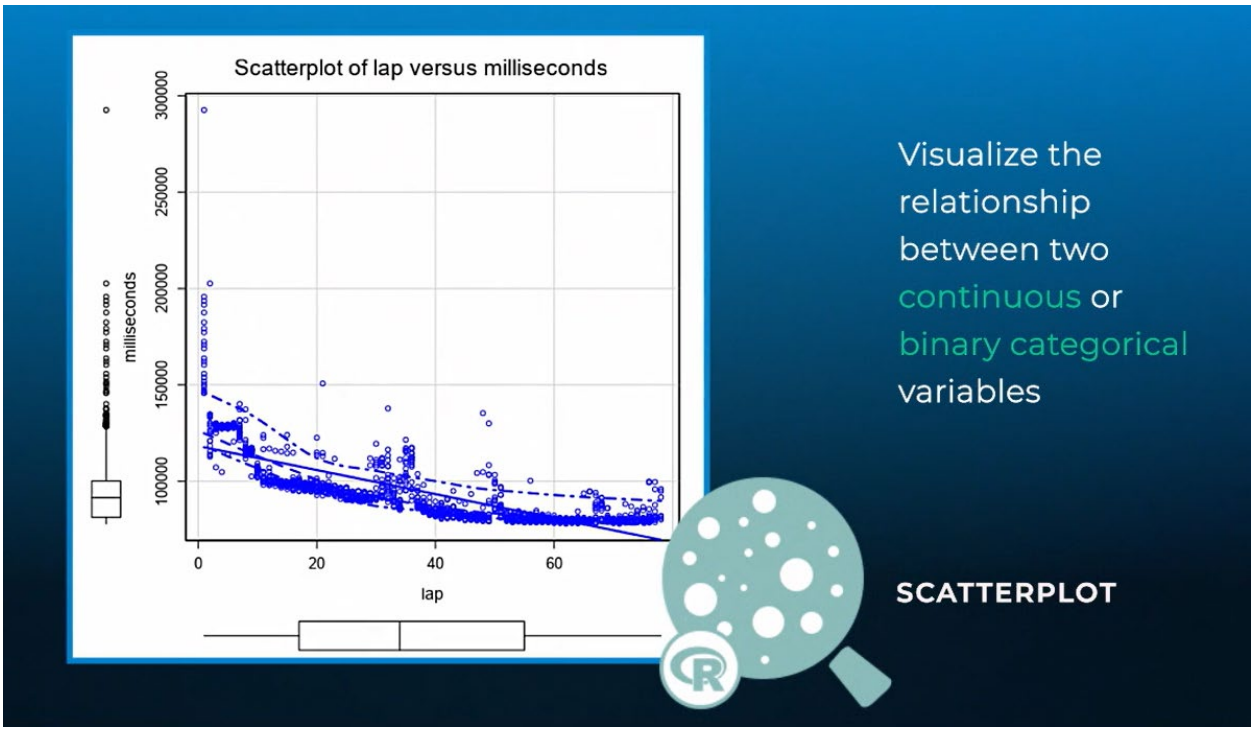
27. Bias & Skew



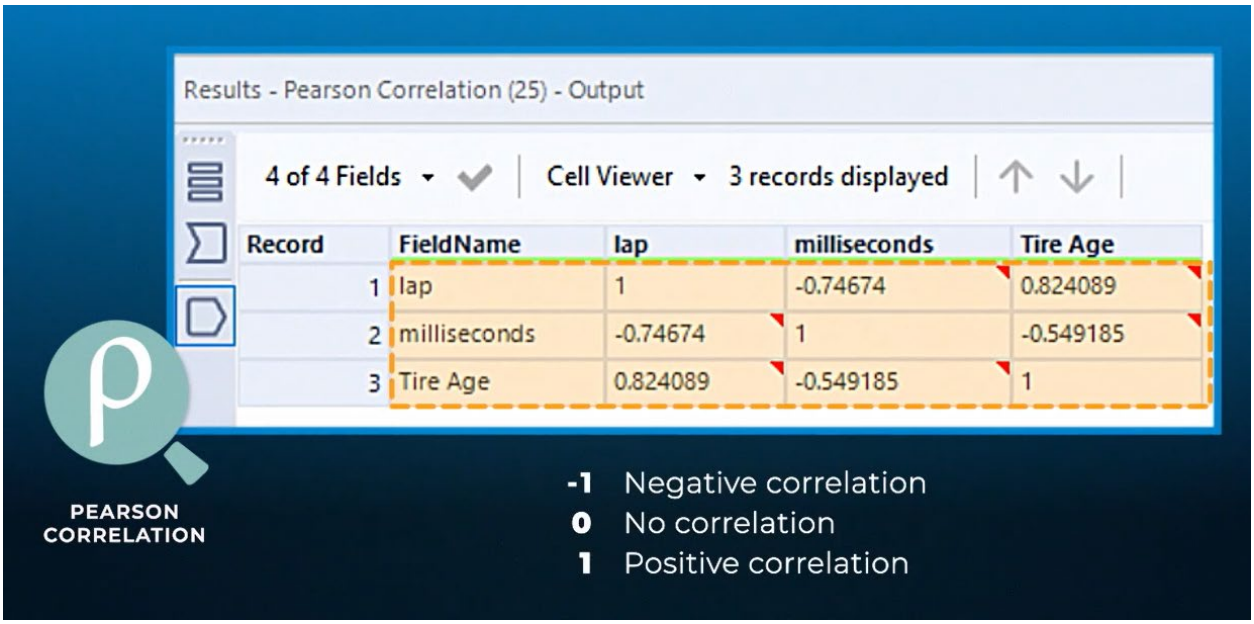
28. Frequency Table



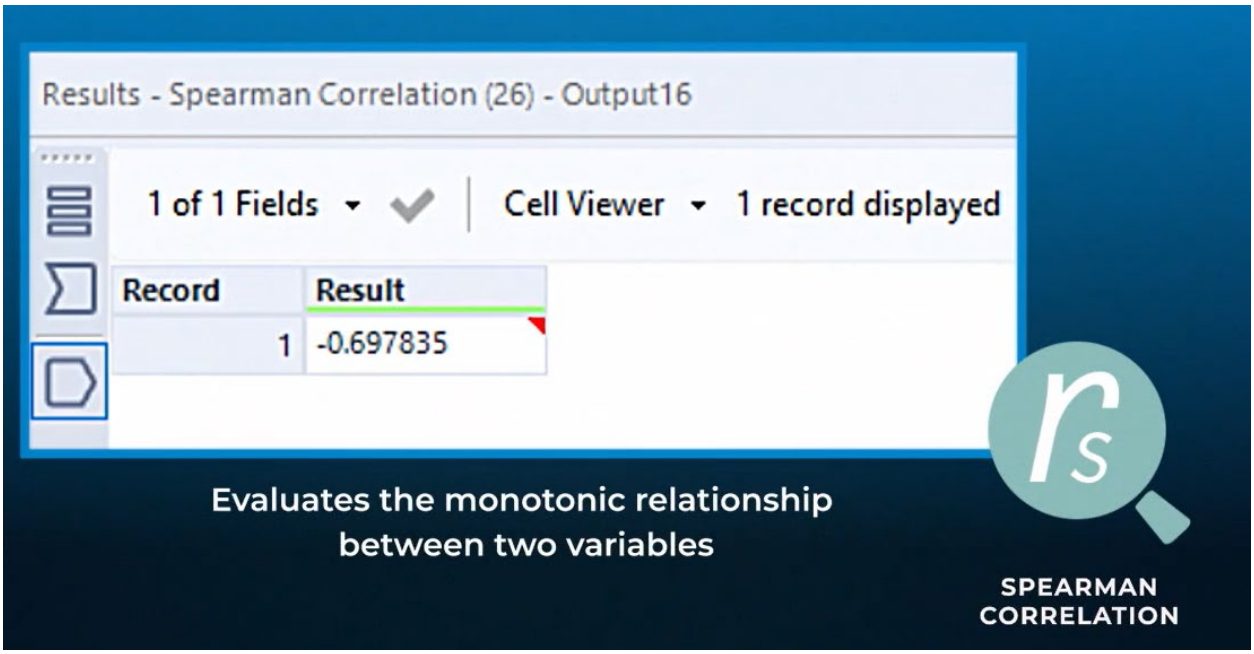
29. Scatterplot

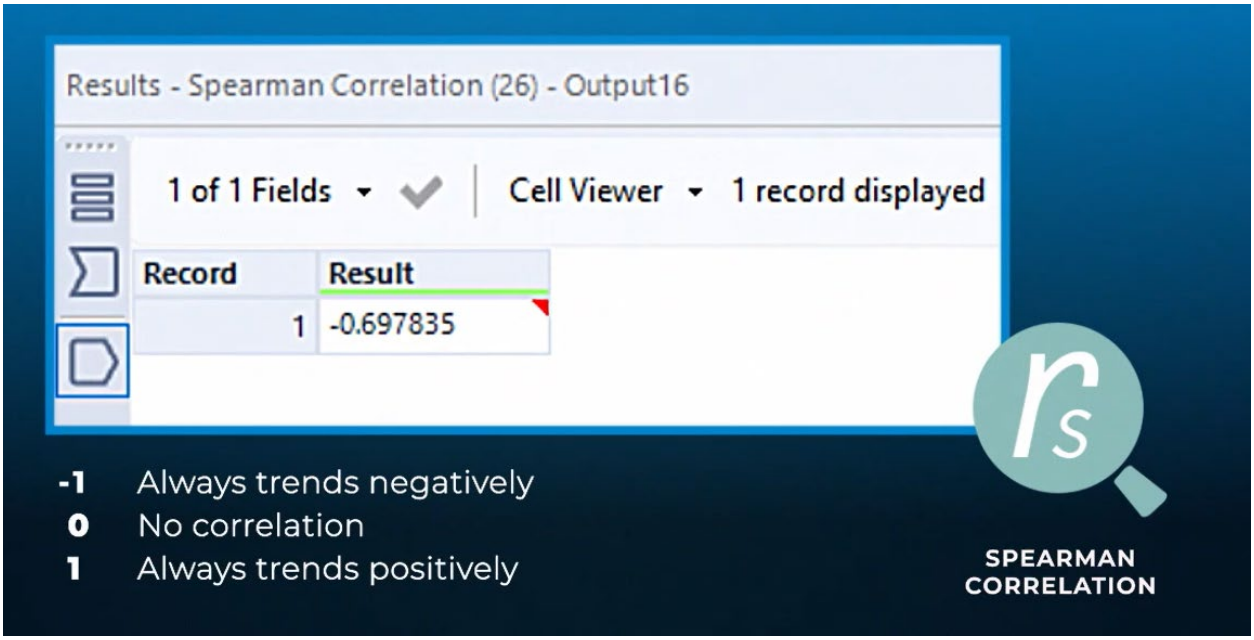


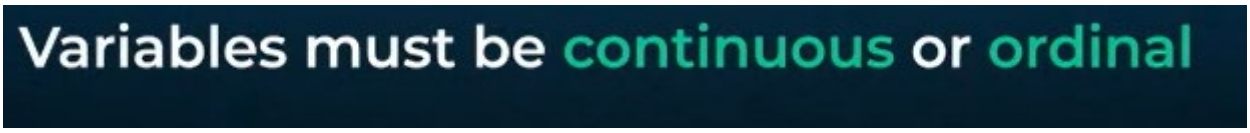
30. Pearson Correlation. Only works for linear relationships.

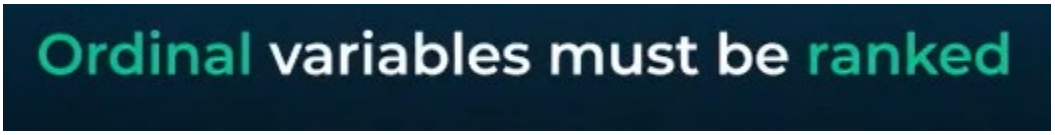


31. Spearman Correlation.

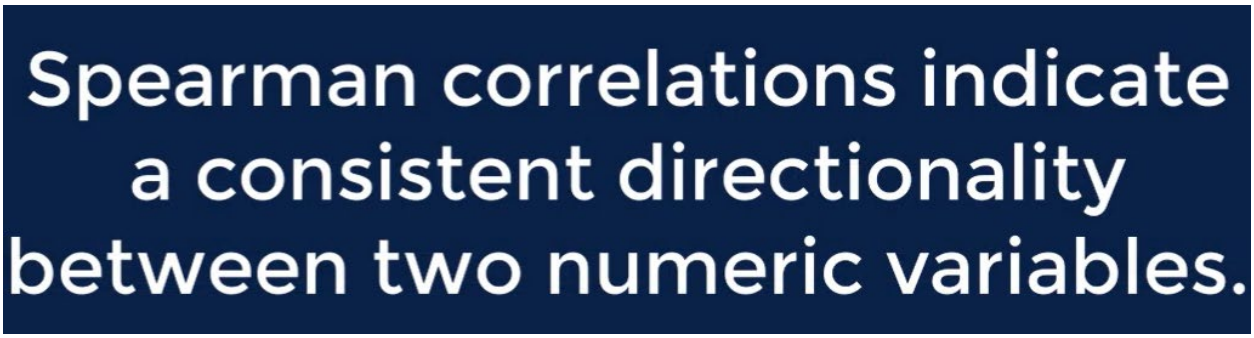




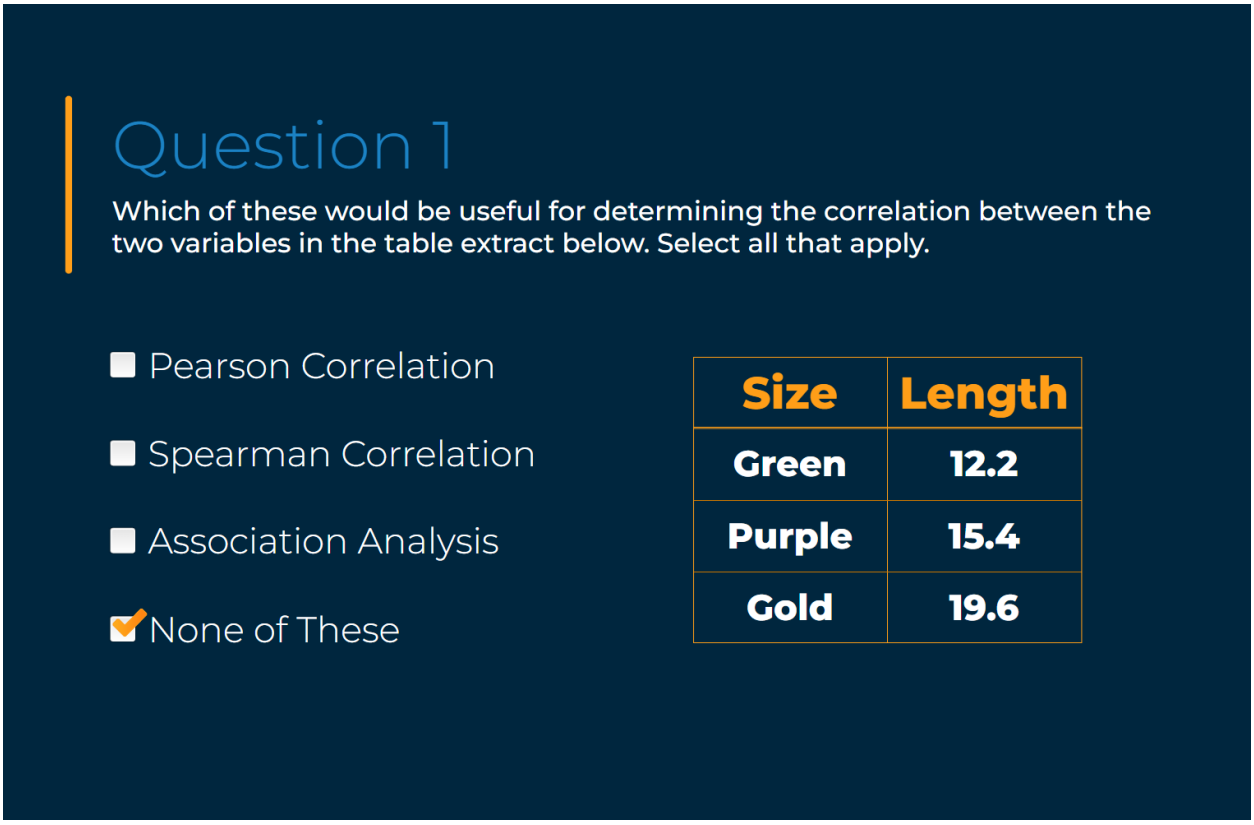




<- Can't say that *red* is a step up from *green*, but *large* IS greater than *small*.



32. Data Investigation question.



33. Data Investigation question.

Question 2

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

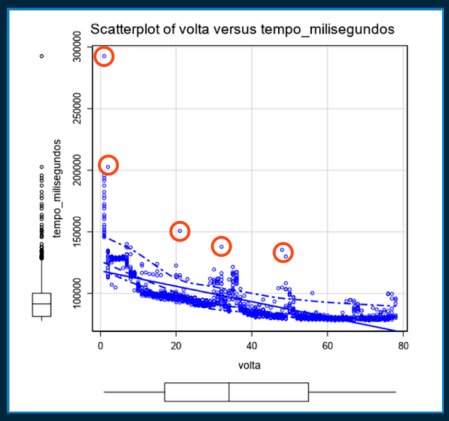
- ☑ String
- ☑ Double
- ☑ DateTime
- ☑ Int16
- ☑ Spatial

34. Data Investigation question.

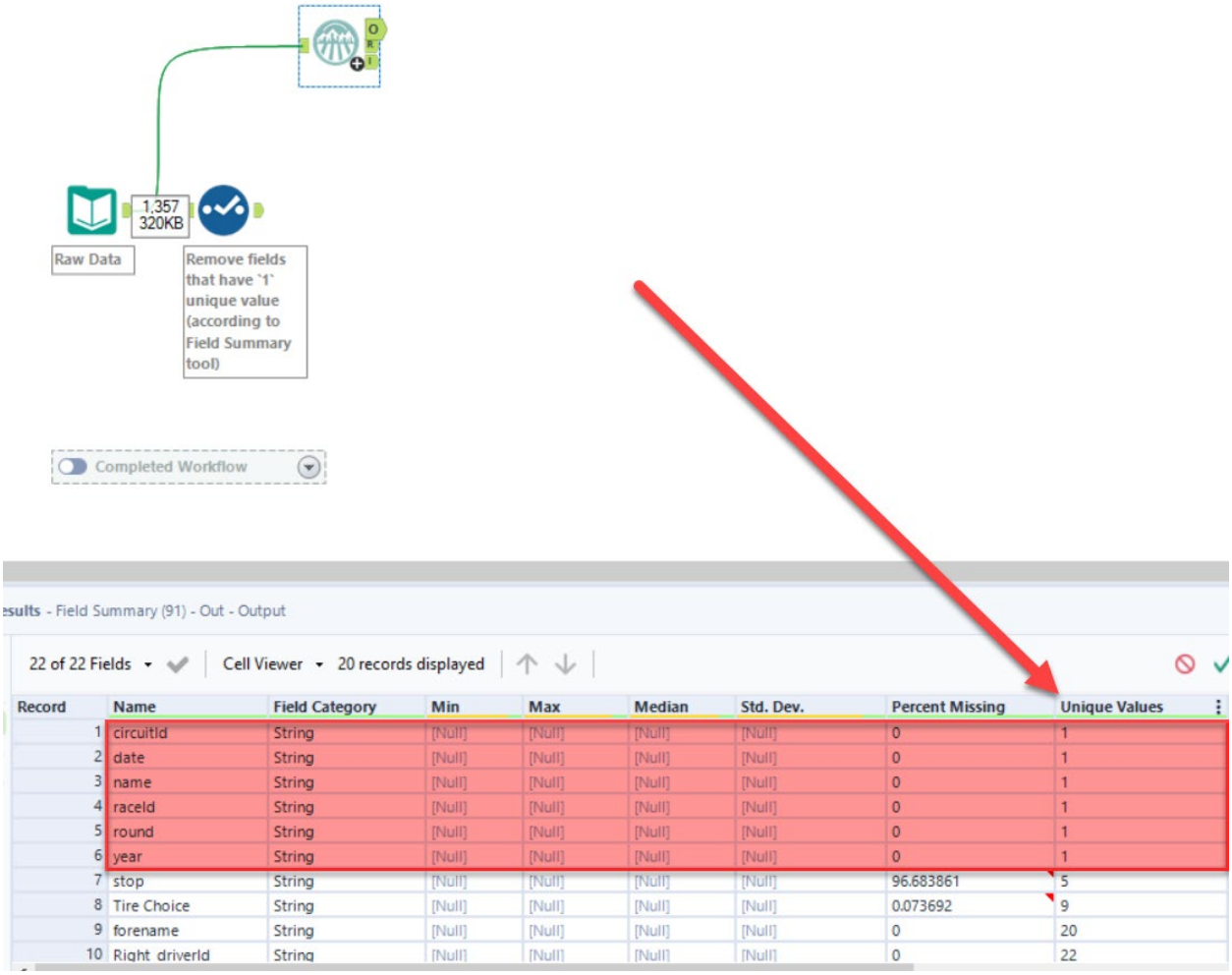
Question 3

What do these dots represent on a scatterplot?

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

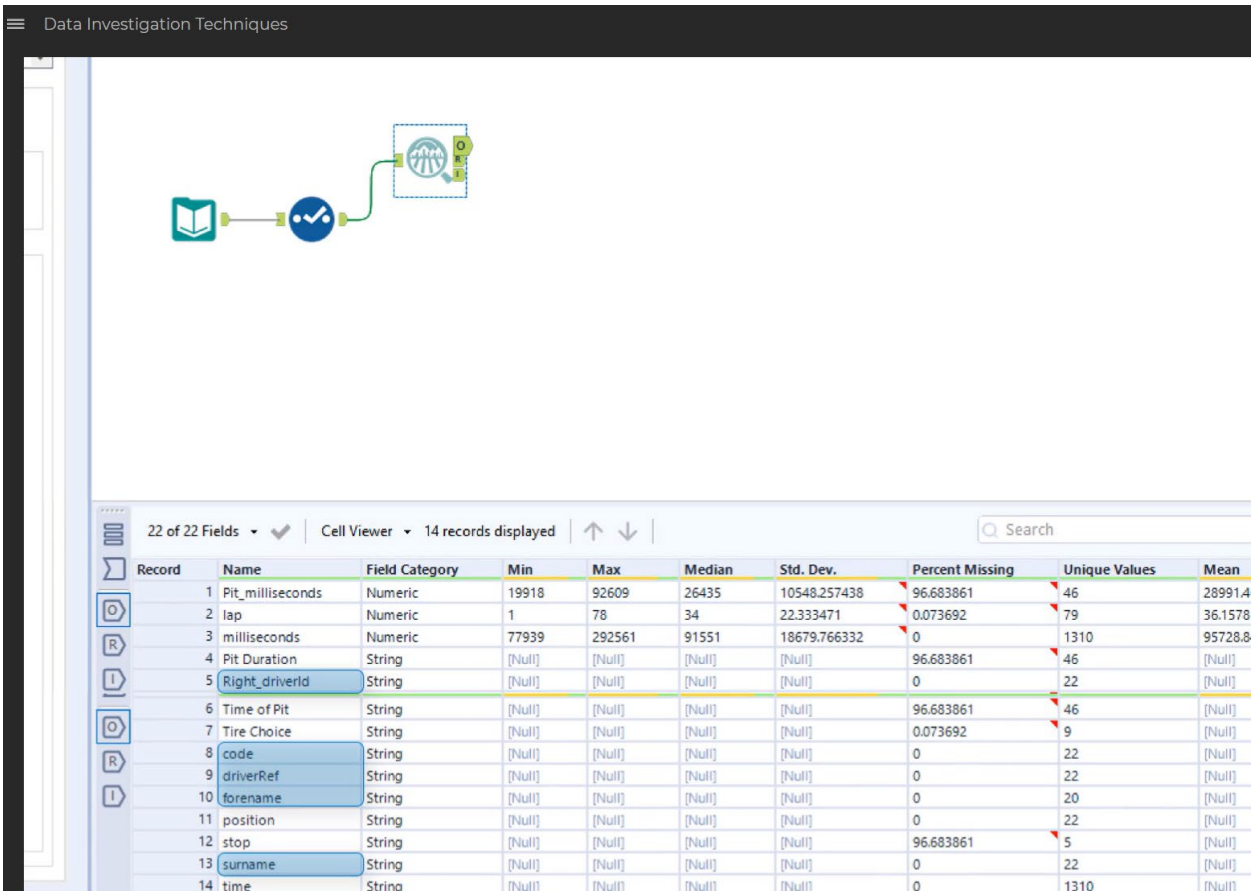


35. Data Investigation: removing fields that have only 1 unique value.



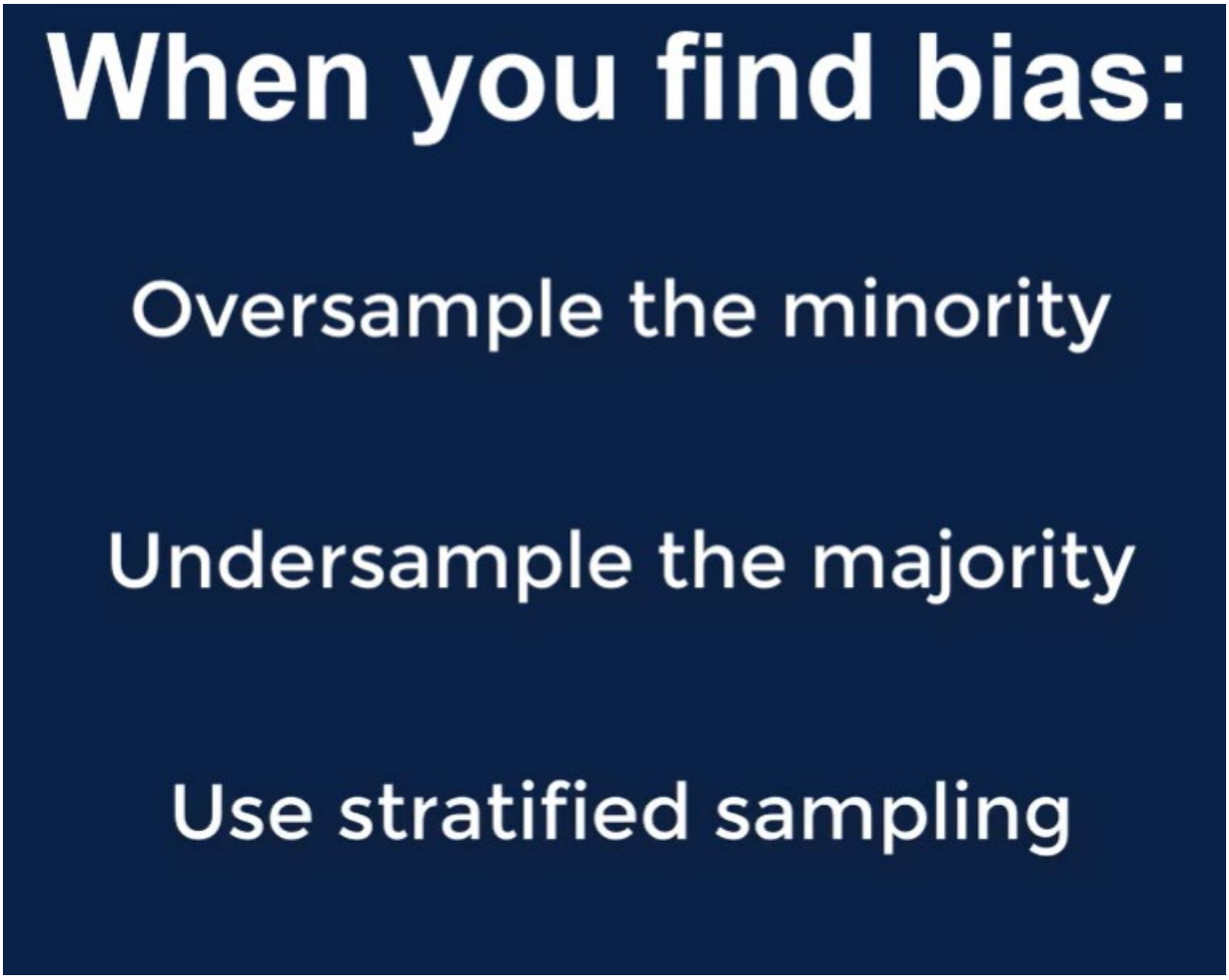
Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values
1	circuitId	String	[Null]	[Null]	[Null]	[Null]	0	1
2	date	String	[Null]	[Null]	[Null]	[Null]	0	1
3	name	String	[Null]	[Null]	[Null]	[Null]	0	1
4	raceId	String	[Null]	[Null]	[Null]	[Null]	0	1
5	round	String	[Null]	[Null]	[Null]	[Null]	0	1
6	year	String	[Null]	[Null]	[Null]	[Null]	0	1
7	stop	String	[Null]	[Null]	[Null]	[Null]	96.683861	5
8	Tire Choice	String	[Null]	[Null]	[Null]	[Null]	0.073692	9
9	forename	String	[Null]	[Null]	[Null]	[Null]	0	20
10	Right driverId	String	[Null]	[Null]	[Null]	[Null]	0	22

36. Data Investigation: removing fields that describe the same attribute (in this example, it's 5 columns that describe the driver). Only the `code` field is kept.

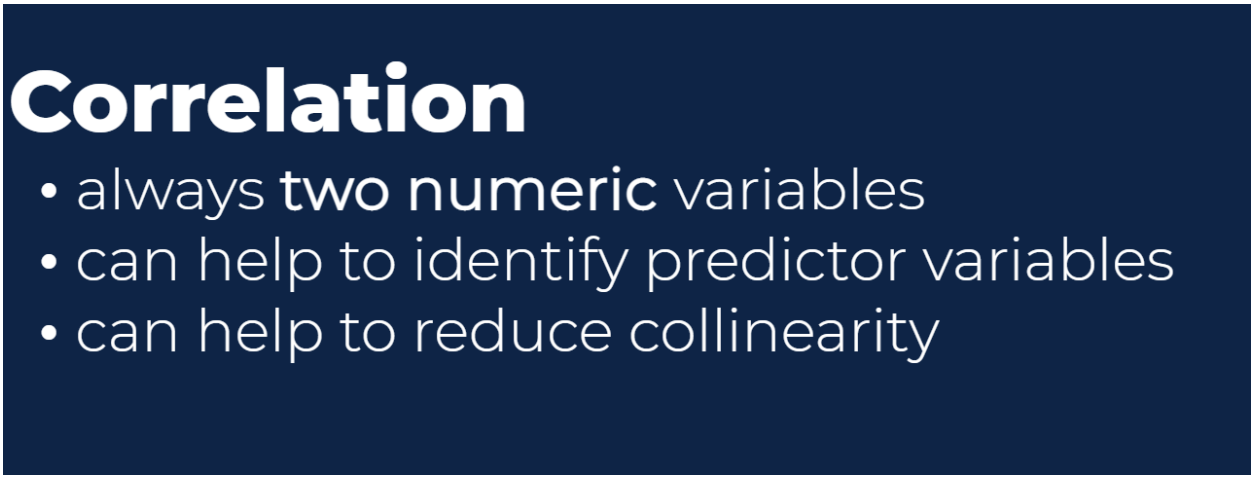


Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	Pit_milliseconds	Numeric	19918	92609	26435	10548.257438	96.683861	46	28991.4
2	lap	Numeric	1	78	34	22.333471	0.073692	79	36.1578
3	milliseconds	Numeric	77939	292561	91551	18679.766332	0	1310	95728.8
4	Pit Duration	String	[Null]	[Null]	[Null]	[Null]	96.683861	46	[Null]
5	Right_driverId	String	[Null]	[Null]	[Null]	[Null]	0	22	[Null]
6	Time of Pit	String	[Null]	[Null]	[Null]	[Null]	96.683861	46	[Null]
7	Tire Choice	String	[Null]	[Null]	[Null]	[Null]	0.073692	9	[Null]
8	code	String	[Null]	[Null]	[Null]	[Null]	0	22	[Null]
9	driverRef	String	[Null]	[Null]	[Null]	[Null]	0	22	[Null]
10	forename	String	[Null]	[Null]	[Null]	[Null]	0	20	[Null]
11	position	String	[Null]	[Null]	[Null]	[Null]	0	22	[Null]
12	stop	String	[Null]	[Null]	[Null]	[Null]	96.683861	5	[Null]
13	surname	String	[Null]	[Null]	[Null]	[Null]	0	22	[Null]
14	time	String	[Null]	[Null]	[Null]	[Null]	0	1310	[Null]

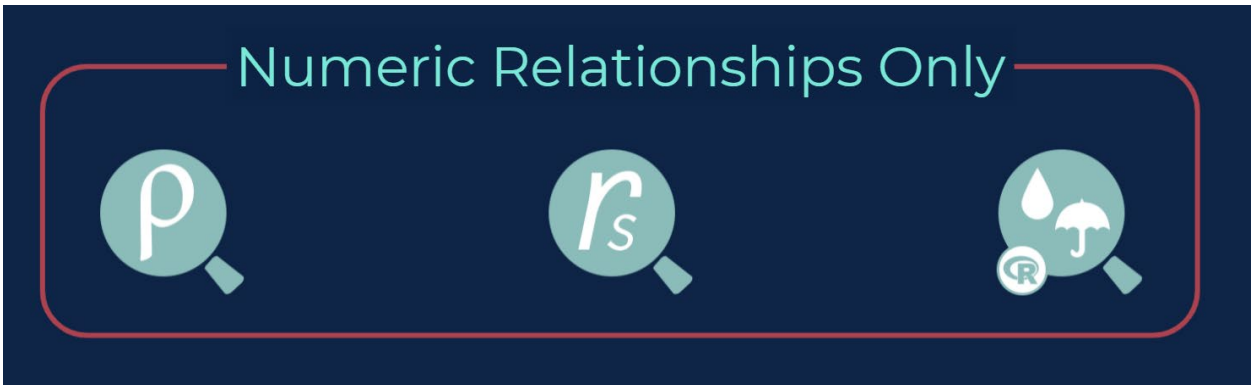
37. When you find bias.



38. Correlation



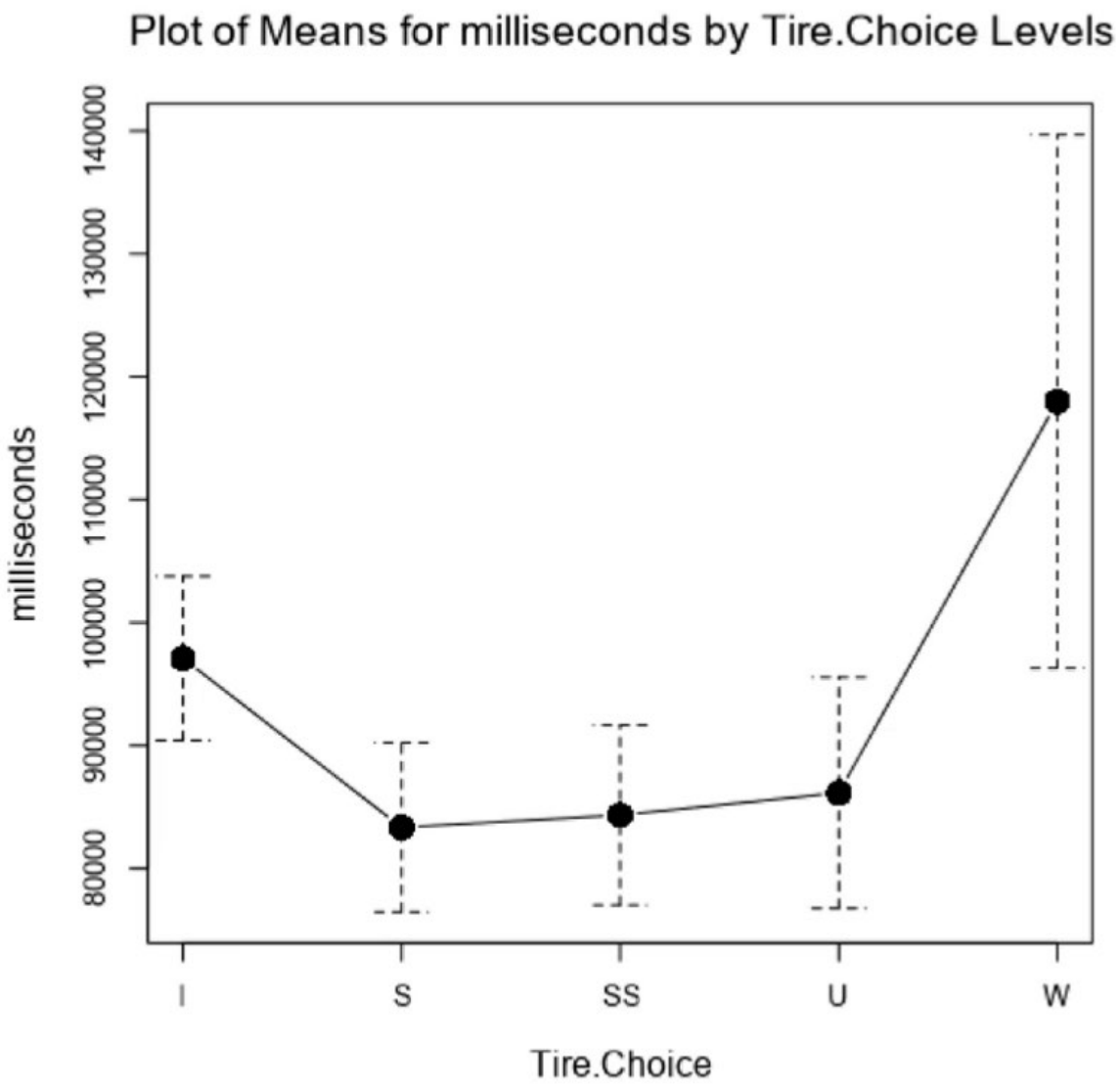
39. These Data Investigation tools can be used to find numeric relationships only.



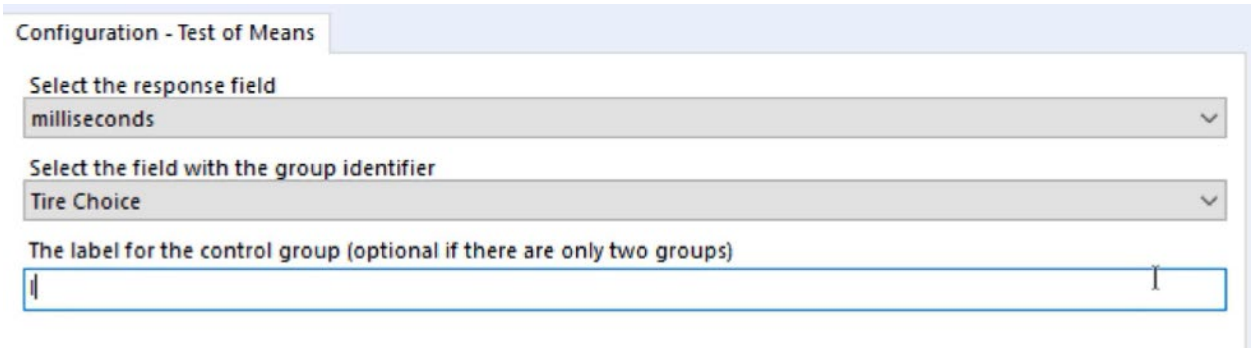
40. These tools (i.e., Plot of Means and Test of Means) can show relationships between continuous and categorical variables.

Example:

Graph



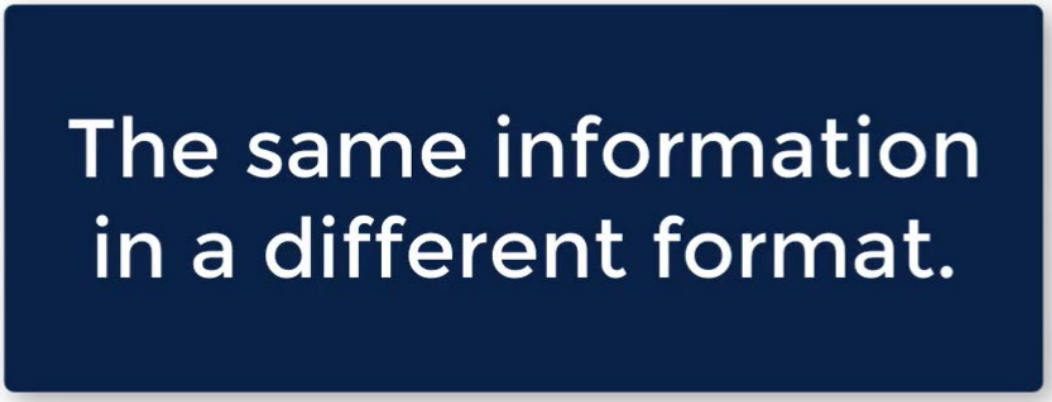
Example:



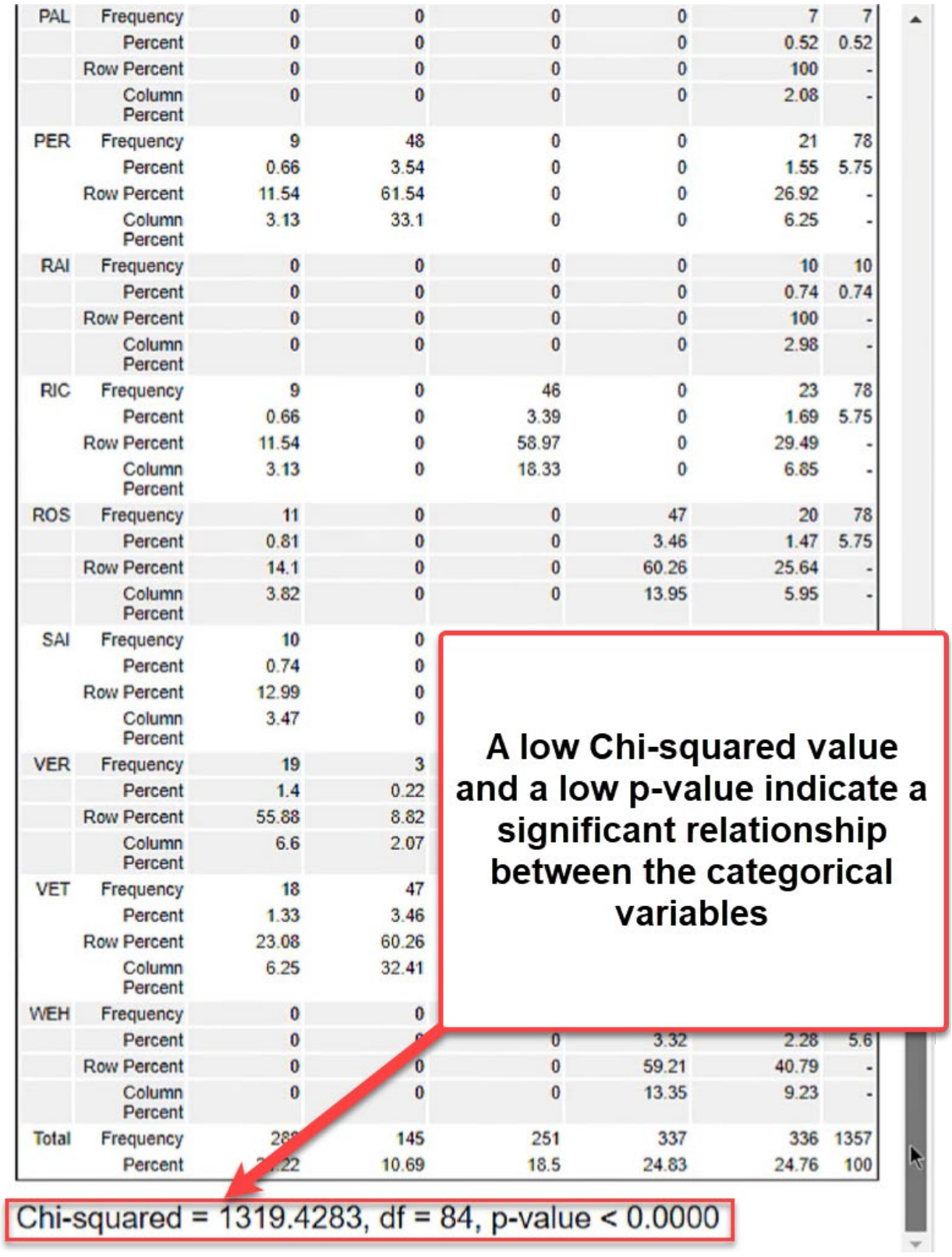
Record	Report
1	**Welch's Two Sample t-test(s) of milliseconds by Tire.Choice**

2

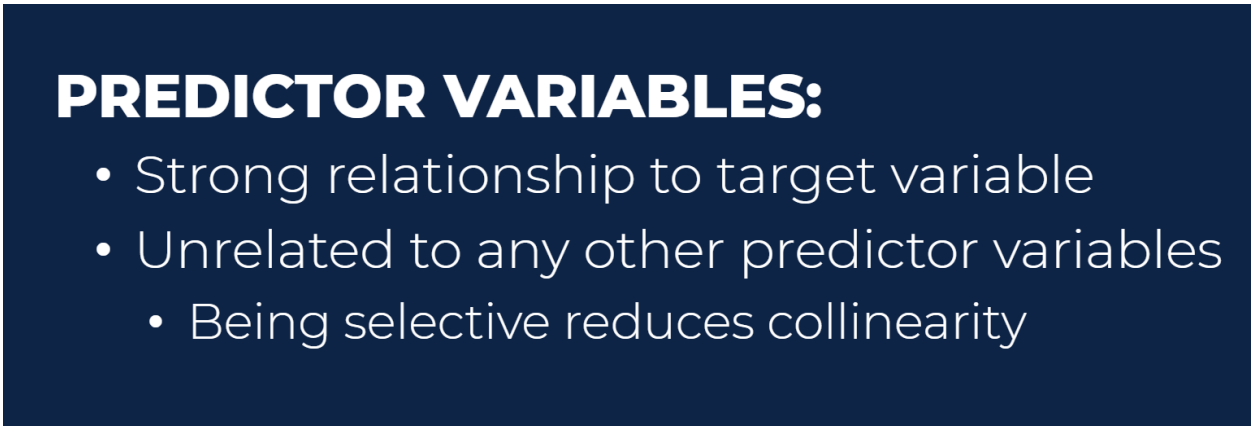
Test	t-Statistic	Degrees of Freedom	p-Value
I vs W	-16.7554	407.45	2.5967e-48
I vs SS	21.015	510.48	4.1459e-71
I vs U	16.9392	603.75	6.0113e-53
I vs S	19.825	281.55	2.3548e-55



41. To test relationships between two categorical variables, the Contingency Table tool can be used.

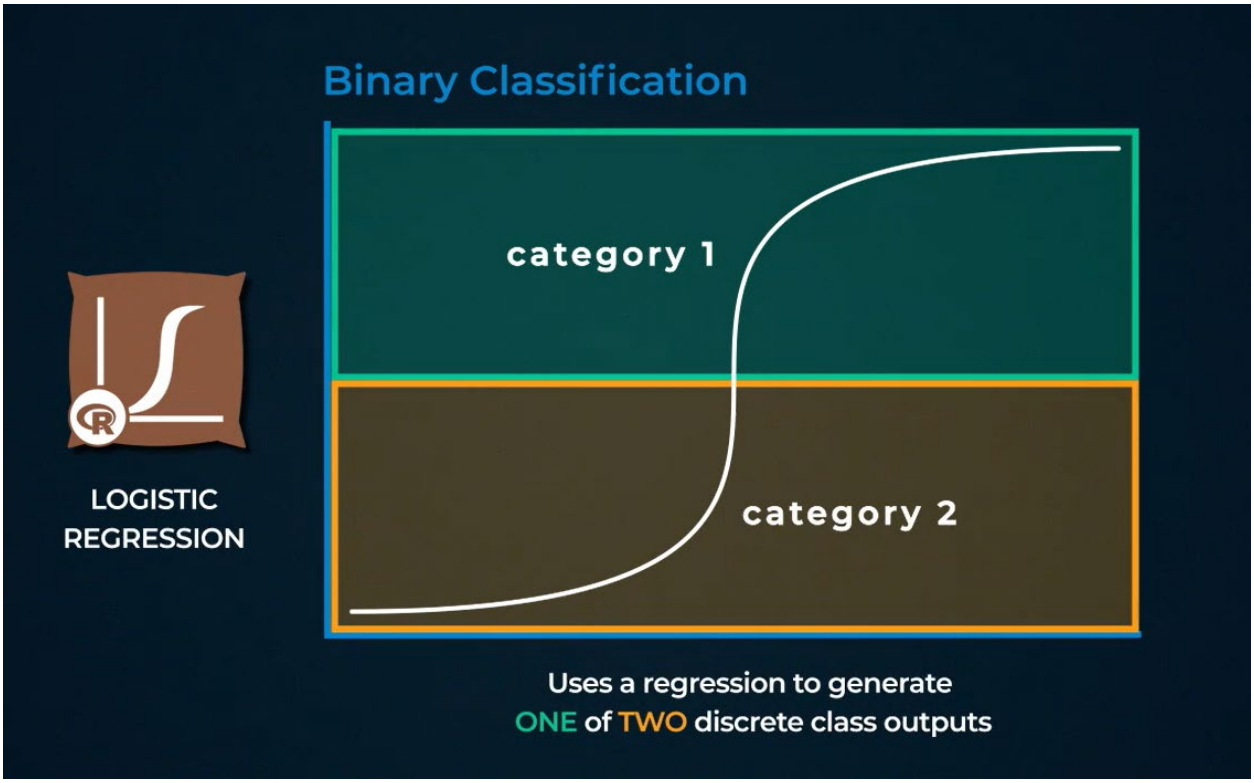


42. Predictor variables.



43. Feature selection

44. Logistic regression



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

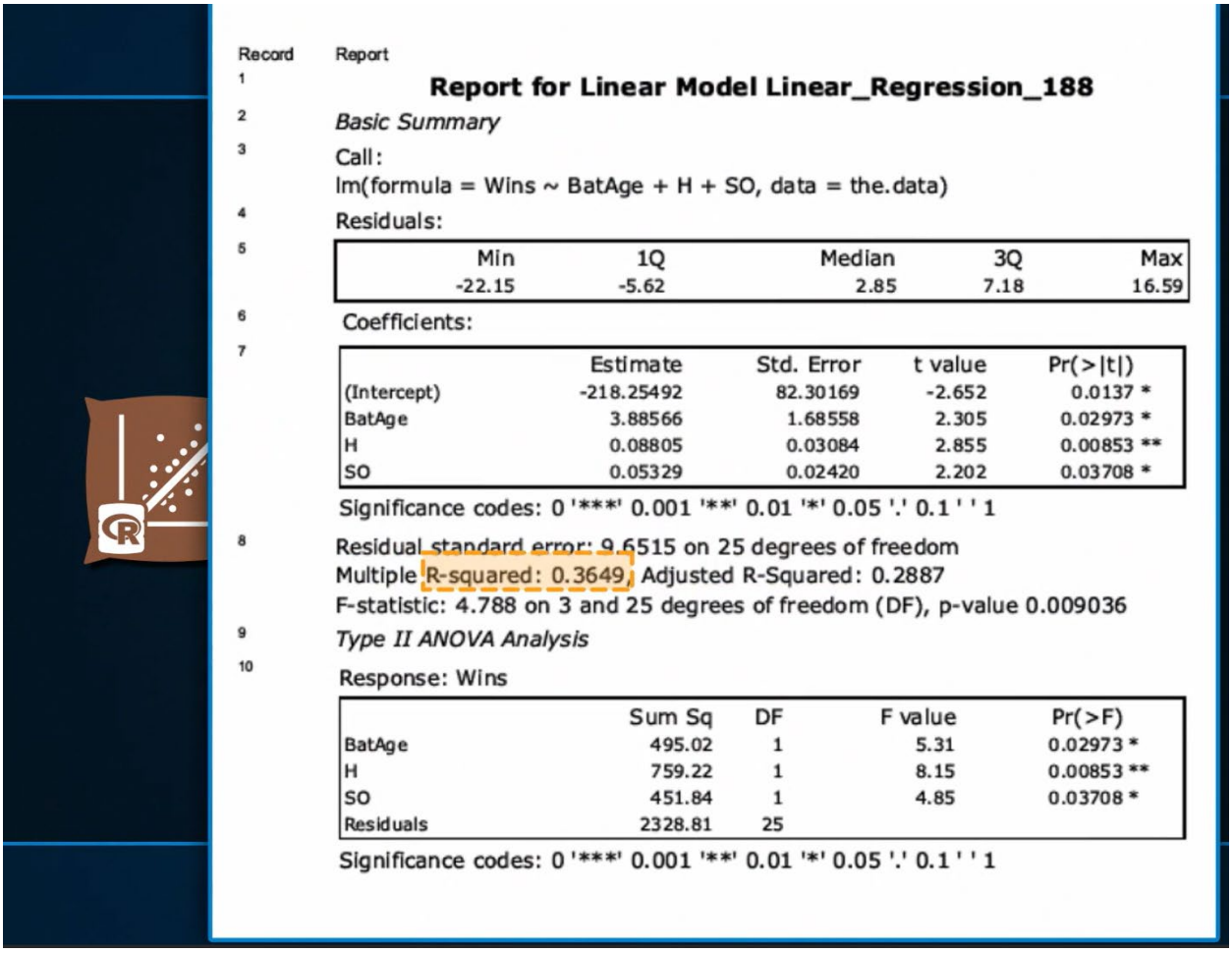
McFadden R-Squared: 0.0648, Akaike Information Criterion 267.4

Number of Fisher Scoring iterations: 4

AIC: Lower is better

Type II Analysis of Deviance Tests

45. Linear Regression. The R-squared statistic ranges from 0 to 1; it measures how much of the variance in the target (i.e., dependent) variable is accounted for by the model. The Adjusted R-squared accounts for the number of variables in the model and can be more representative of the model's performance.



Record Report

Report for Linear Model Linear_Regression_188

Basic Summary

Call:

lm(formula = Wins ~ BatAge + H + SO, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-22.15	-5.62	2.85	7.18	16.59

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-218.25492	82.30169	-2.652	0.0137 *
BatAge	3.88566	1.68558	2.305	0.02973 *
H	0.08805	0.03084	2.855	0.00853 **
SO	0.05329	0.02420	2.202	0.03708 *

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

Residual standard error: 9.6515 on 25 degrees of freedom

Multiple R-squared: 0.3649, Adjusted R-Squared: 0.2887

F-statistic: 4.788 on 3 and 25 degrees of freedom (DF), p-value 0.009036

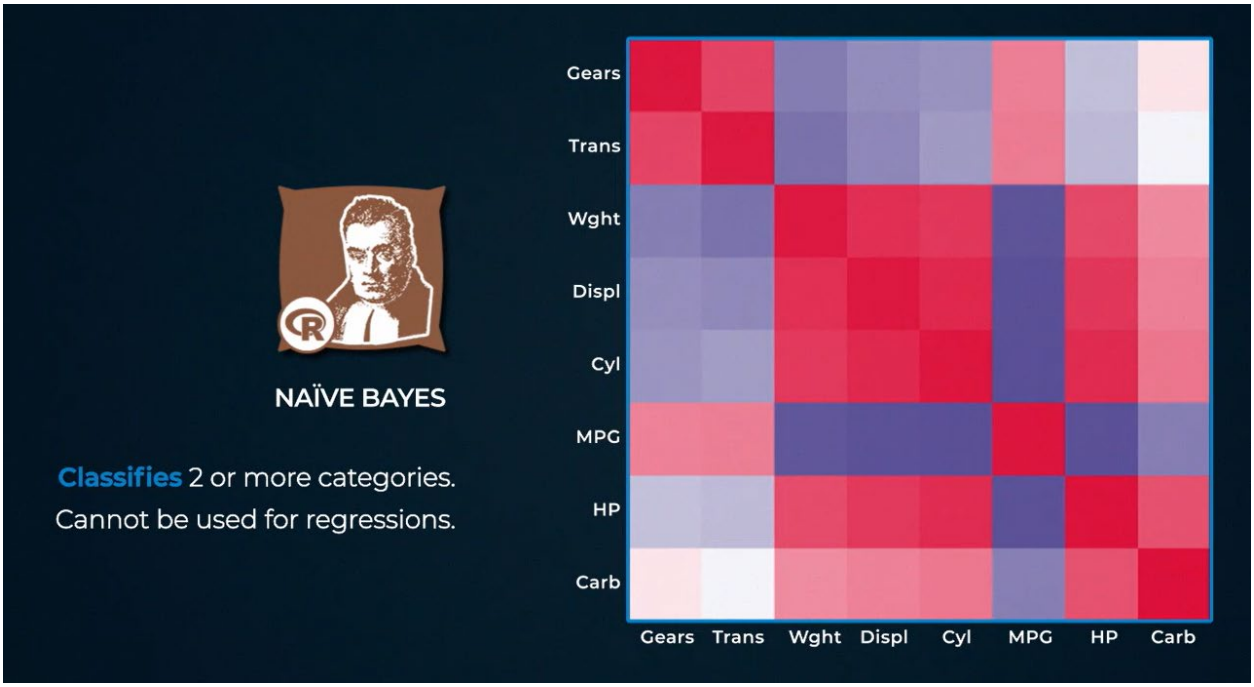
Type II ANOVA Analysis

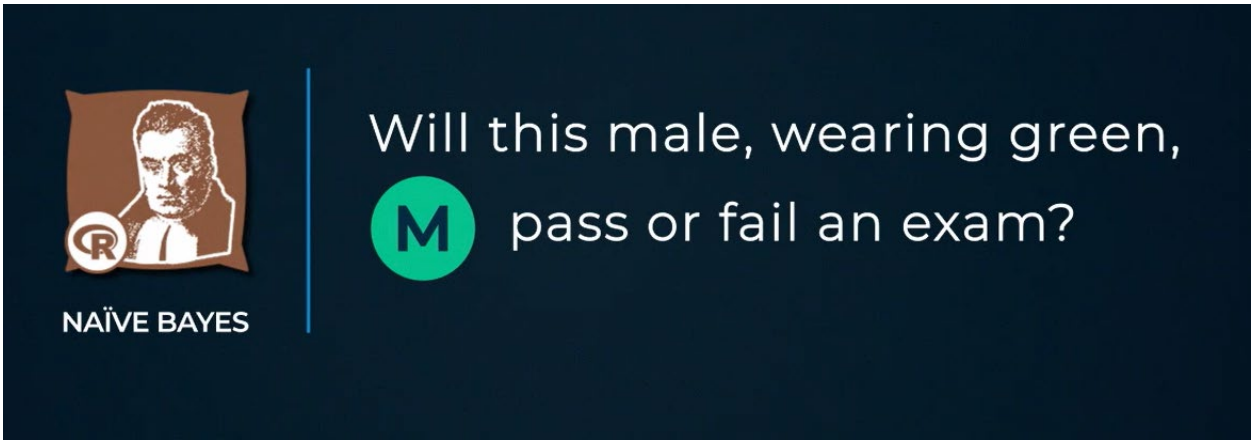
Response: Wins

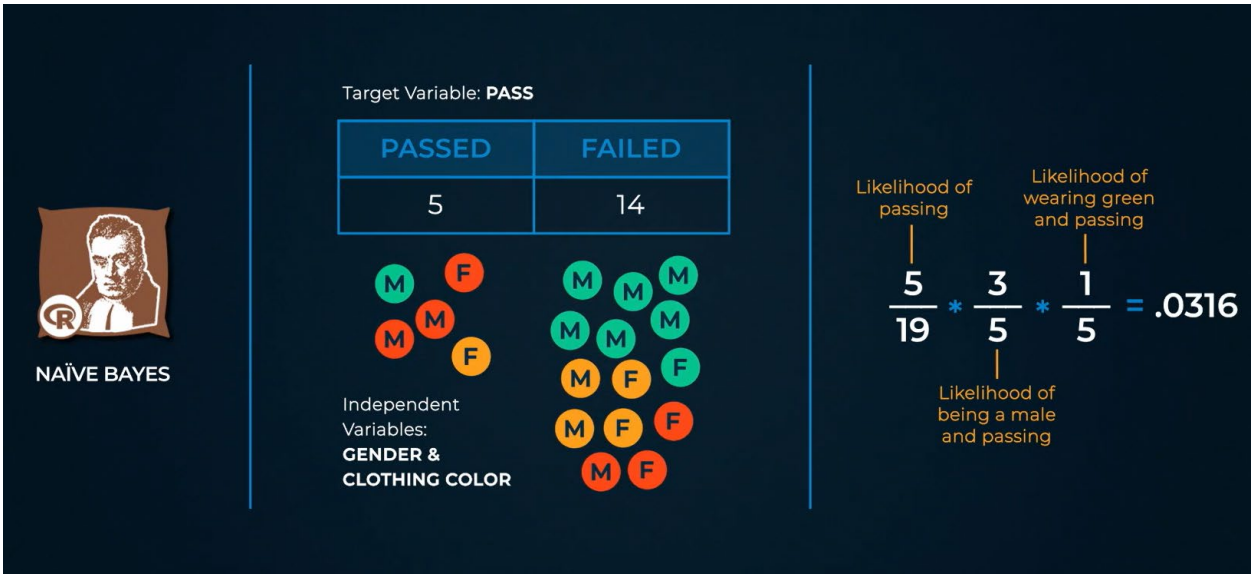
	Sum Sq	DF	F value	Pr(>F)
BatAge	495.02	1	5.31	0.02973 *
H	759.22	1	8.15	0.00853 **
SO	451.84	1	4.85	0.03708 *
Residuals	2328.81	25		

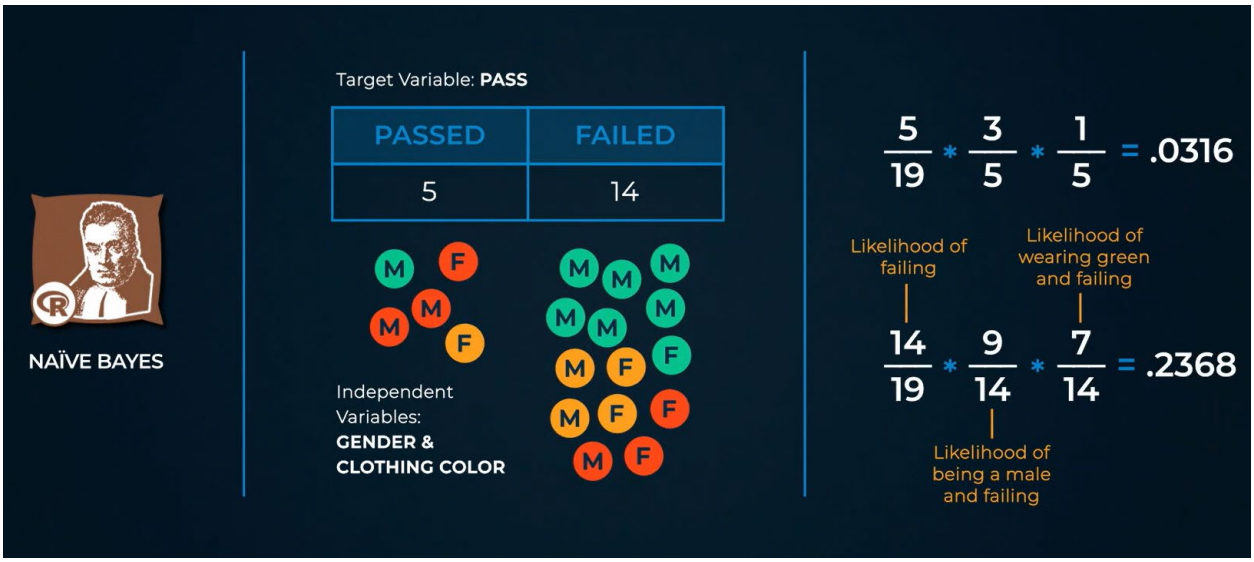
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

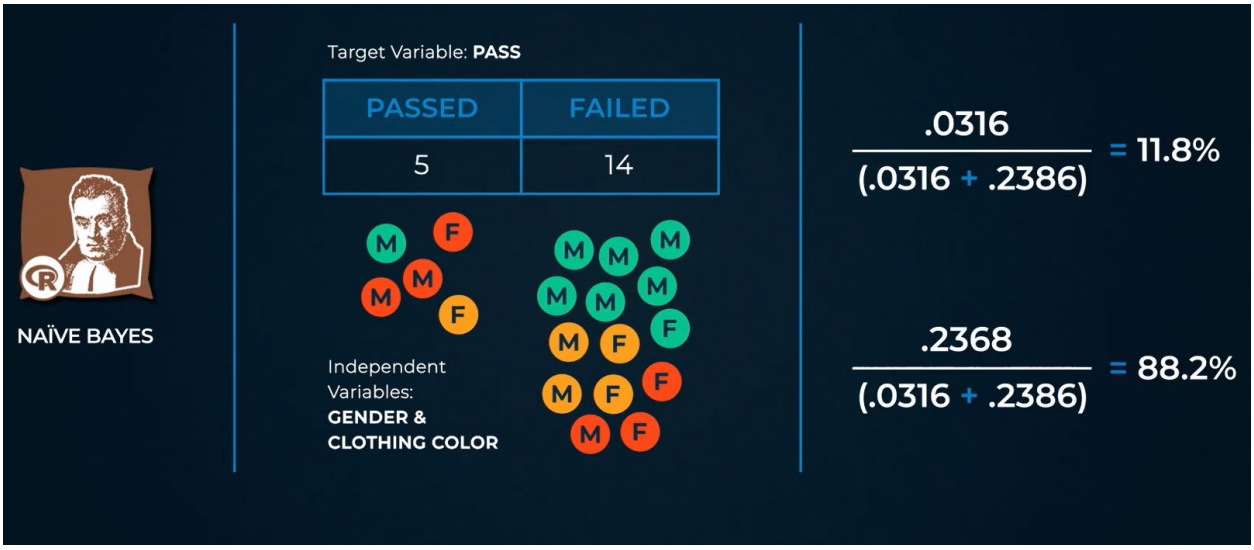
46. Naïve Bayes. Called naïve because it assumes all predictor (i.e., independent) variables are totally independent.


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

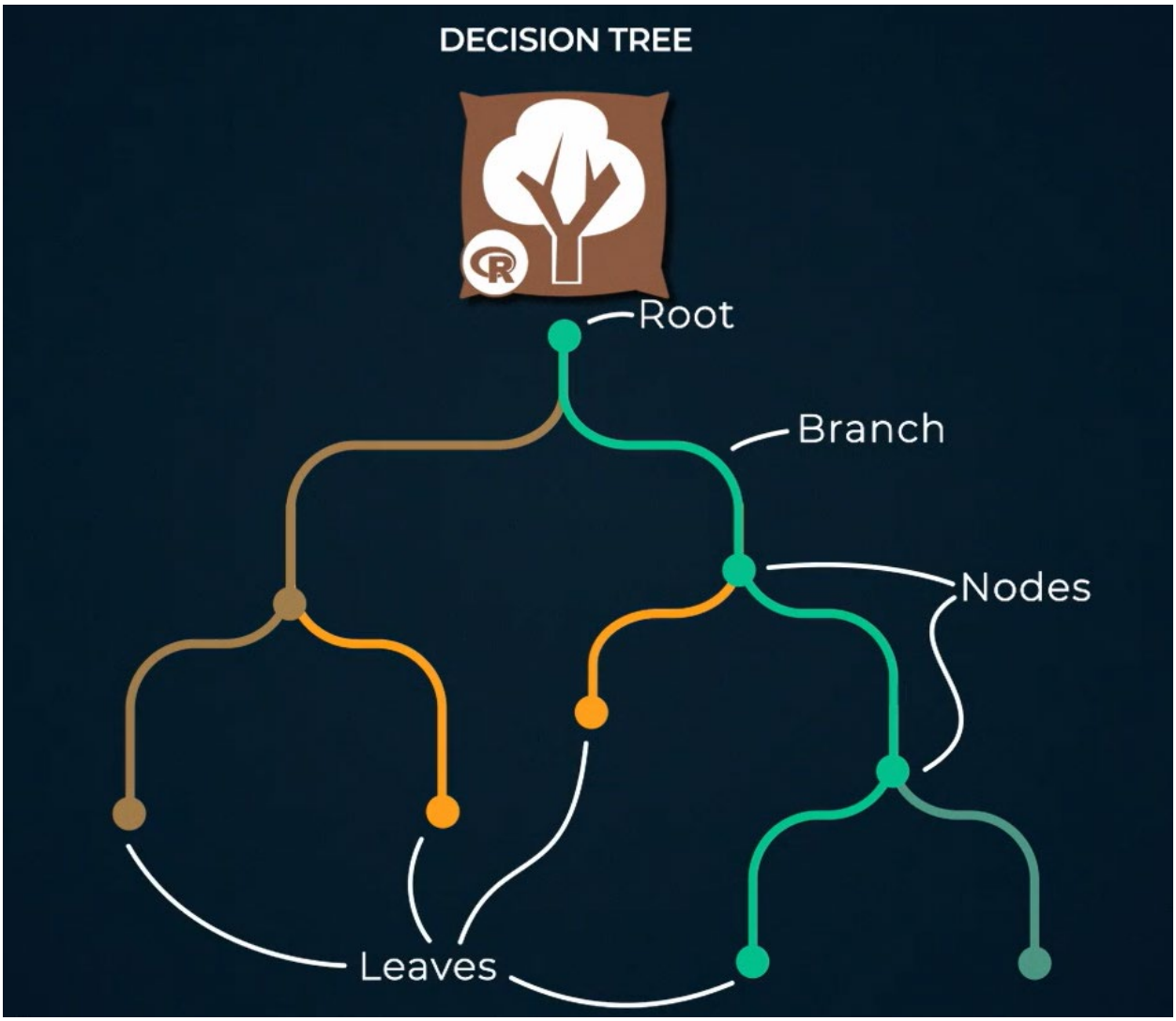

NAÏVE BAYES
Will this male, wearing green,
M
pass or fail an exam?


NAÏVE BAYES
Target Variable: PASS
PASSED
FAILED
5
14
Independent Variables: GENDER & CLOTHING COLOR
Likelihood of passing
Likelihood of wearing green and passing
Likelihood of being a male and passing
5/19 * 3/5 * 1/5 = .0316


NAÏVE BAYES
Target Variable: PASS
PASSED
FAILED
5
14
Independent Variables: GENDER & CLOTHING COLOR
5/19 * 3/5 * 1/5 = .0316
Likelihood of failing
Likelihood of wearing green and failing
Likelihood of being a male and failing
14/19 * 9/14 * 7/14 = .2368


NAÏVE BAYES
Target Variable: PASS
PASSED
FAILED
5
14
Independent Variables: GENDER & CLOTHING COLOR
.0316 / (.0316 + .2386) = 11.8%
.2368 / (.0316 + .2386) = 88.2%

47. Decision Tree



DECISION TREE

PROS

- Easy to Visualize & Interpret
- Easy to Share Results
- Robust to Outliers
- Self-select most important variables

CONS

- Prone to Overfitting
- Sensitivie to changes in training data
- Regression trees result in jagged values

DECISION TREE

>9 = ▭ >8.4 = ▭

<9 = ▭ <8.4 = ▭

SUBJECT	CATEGORY 1	CATEGORY 2	TARGET
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	●

Gini Impurity / Entropy:

- A metric of the split quality

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)

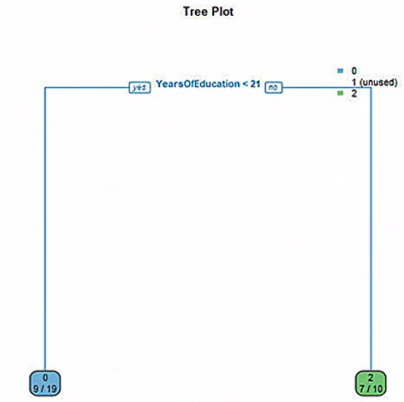
* denotes terminal node

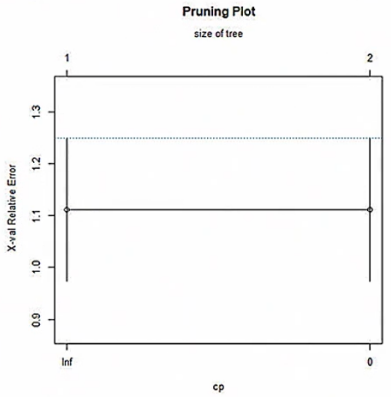
1) root 29 18 0 (0.37931034 0.34482759 0.27586207)

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) *

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





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

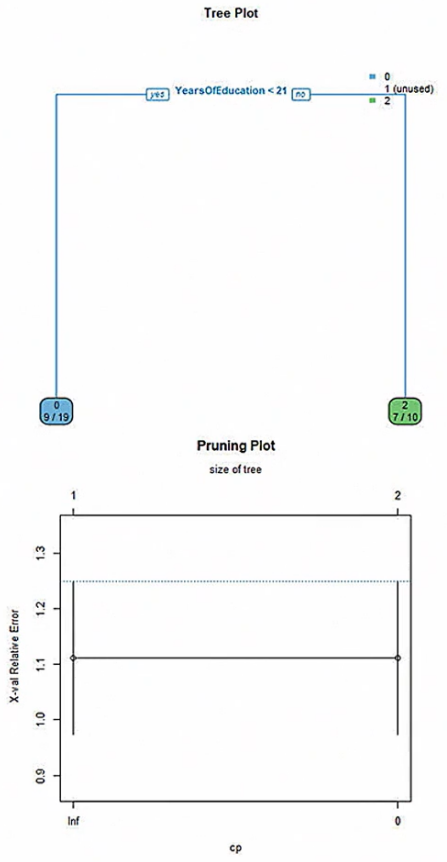
1) root 29 18 0 (0.37931034 0.34482759 0.27586207)

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

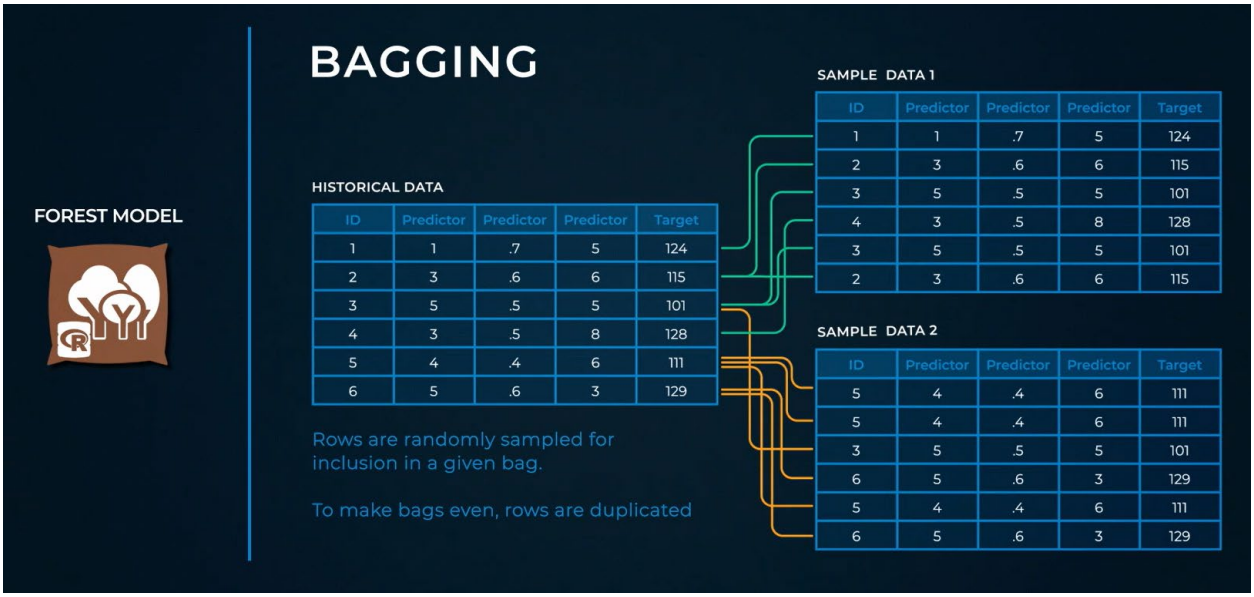
3) YearsOfEducation>=20.5 10 3 2 (0.20000000 0.10000000 0.70000000) *

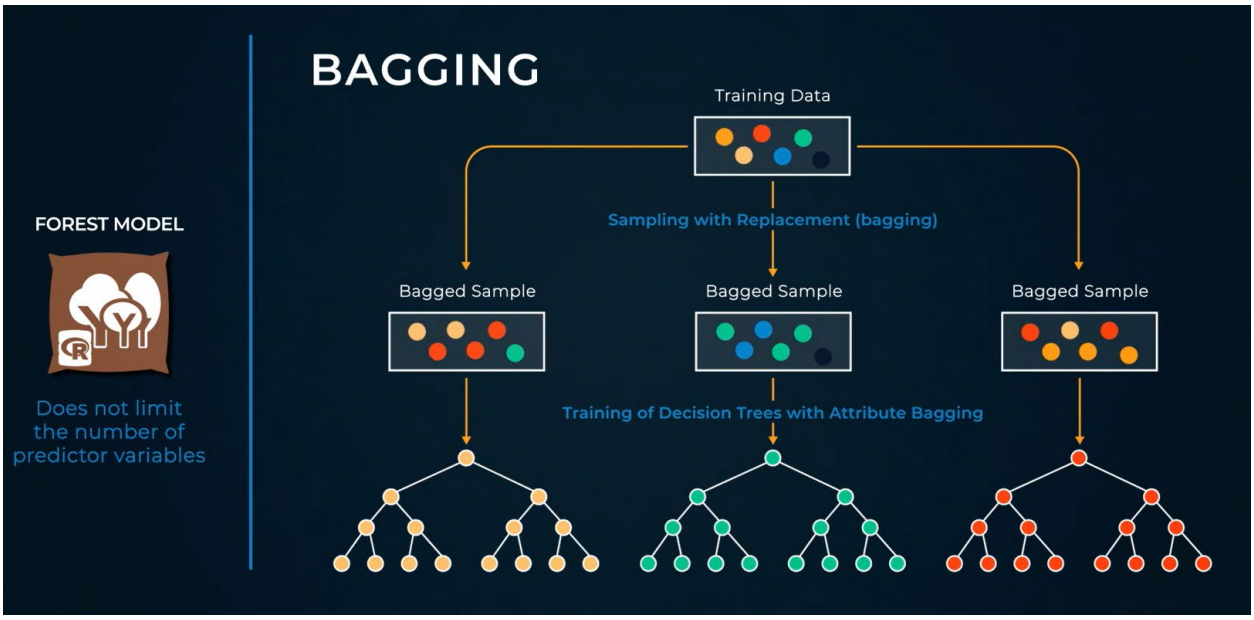
Add these values and compare to the xError.

If sum > xError, then prune that branch.

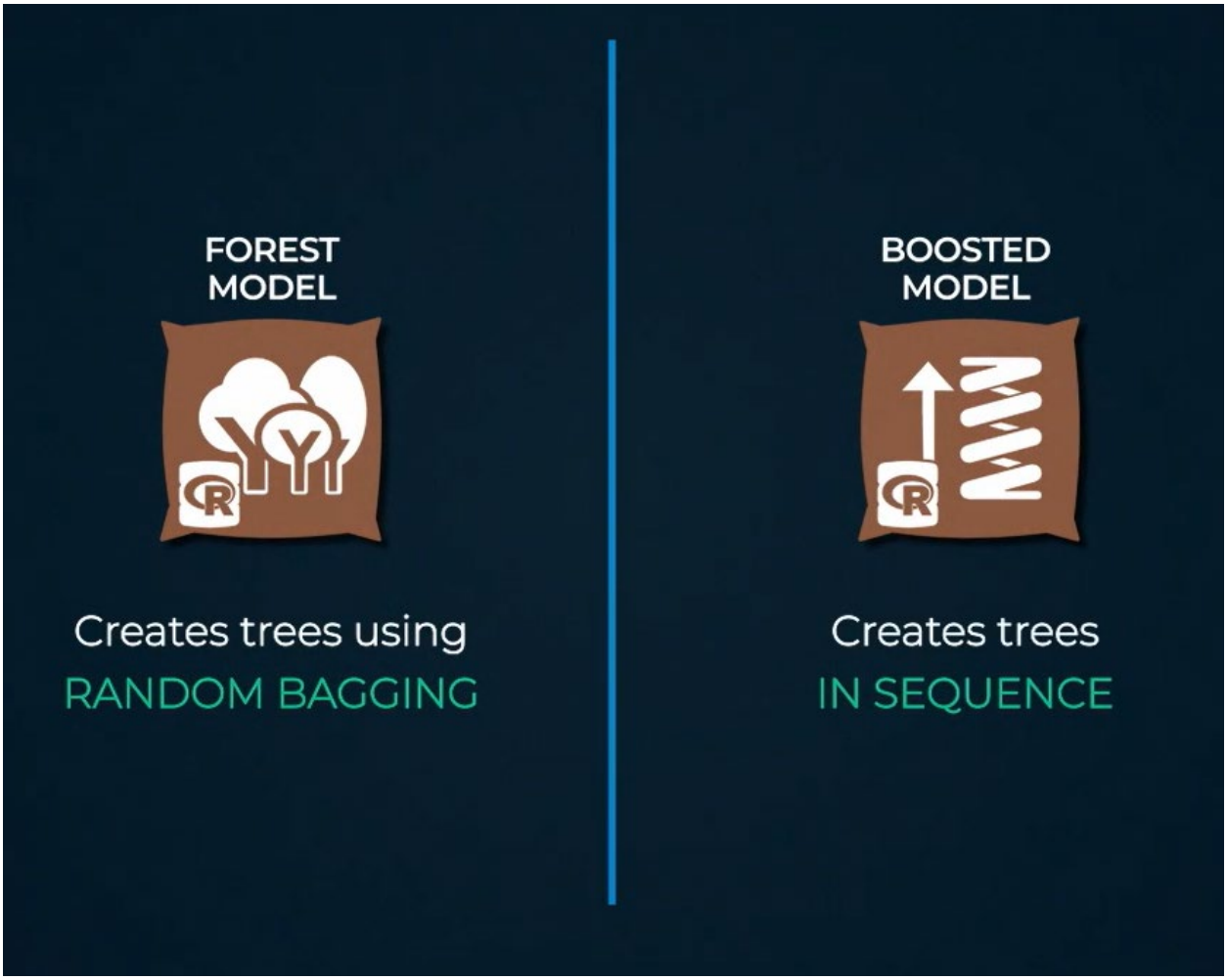


48. Forest model





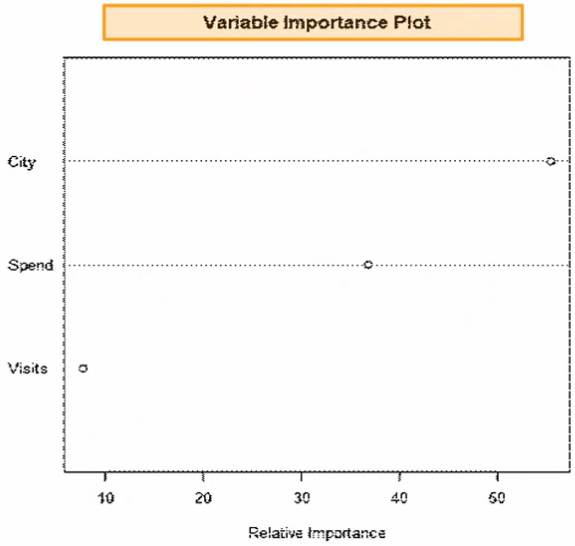
49. Boosted model

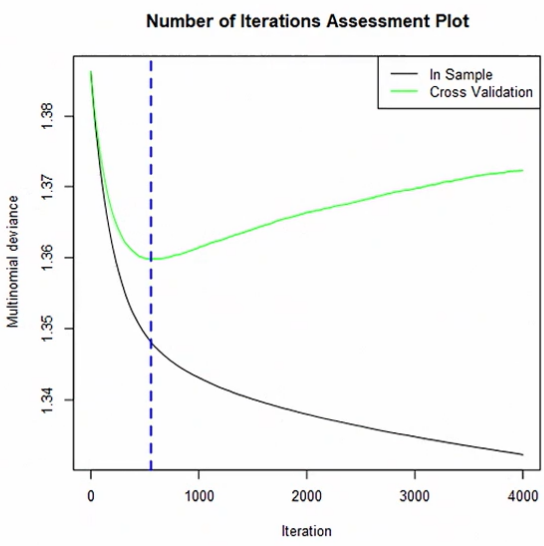

FOREST
MODEL
Creates trees using
RANDOM BAGGING
BOOSTED
MODEL
Creates trees
IN SEQUENCE

Report for Boosted Model Boosted

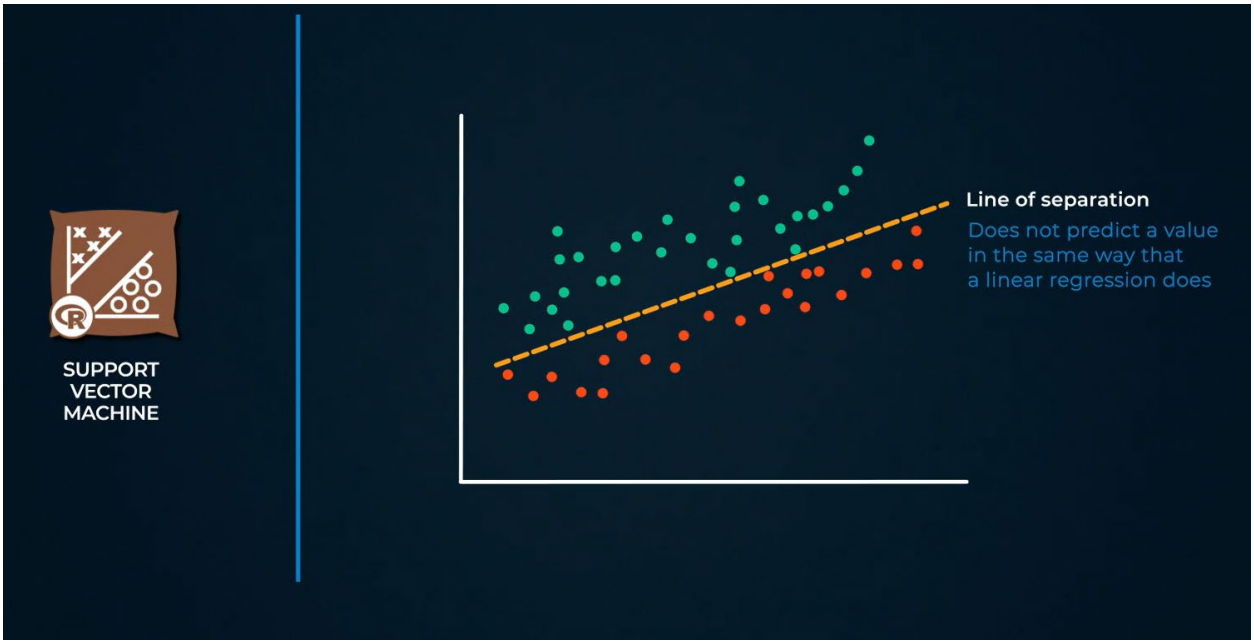
Basic Summary:

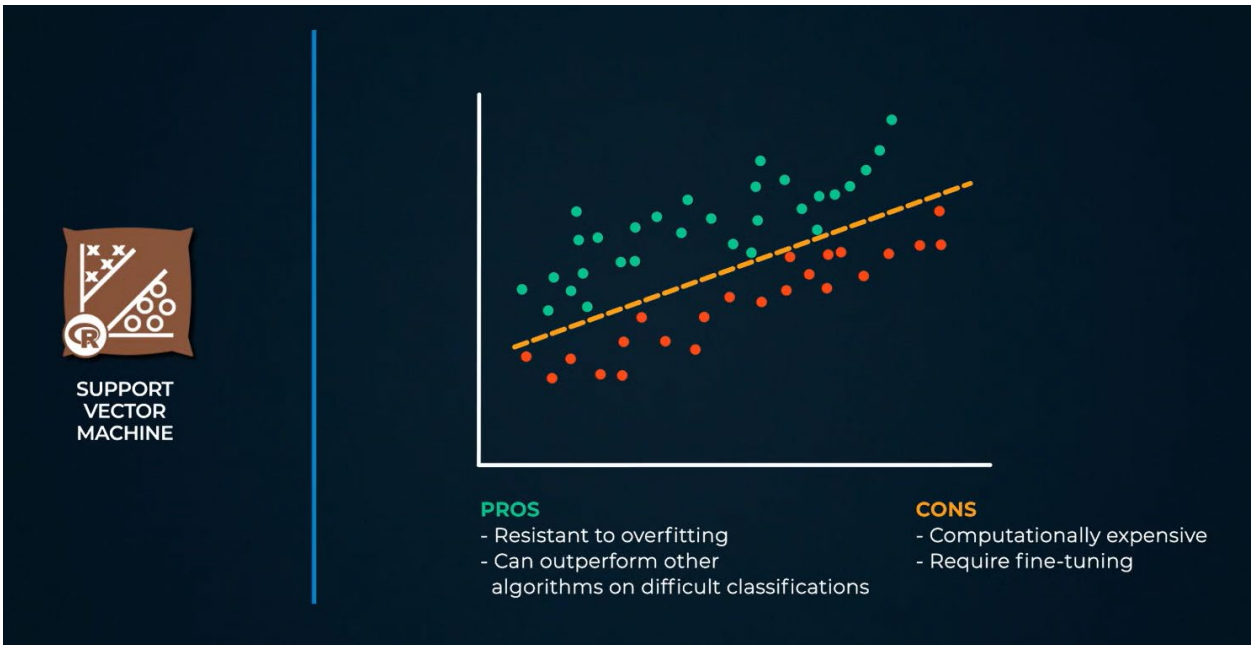
Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 557

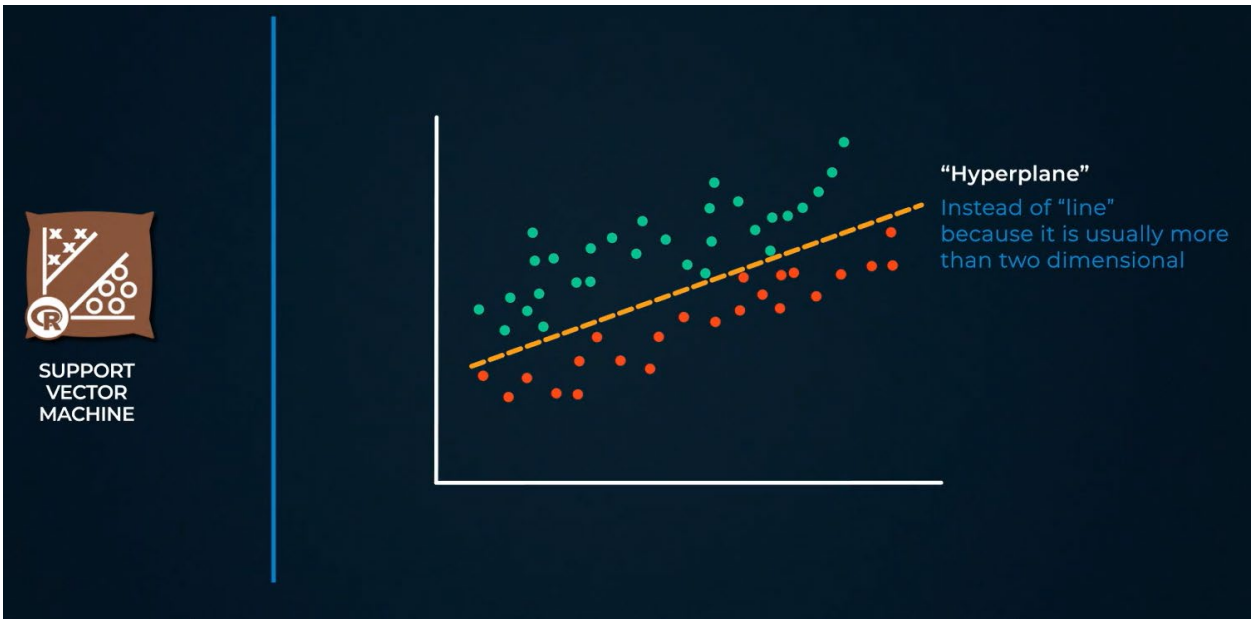


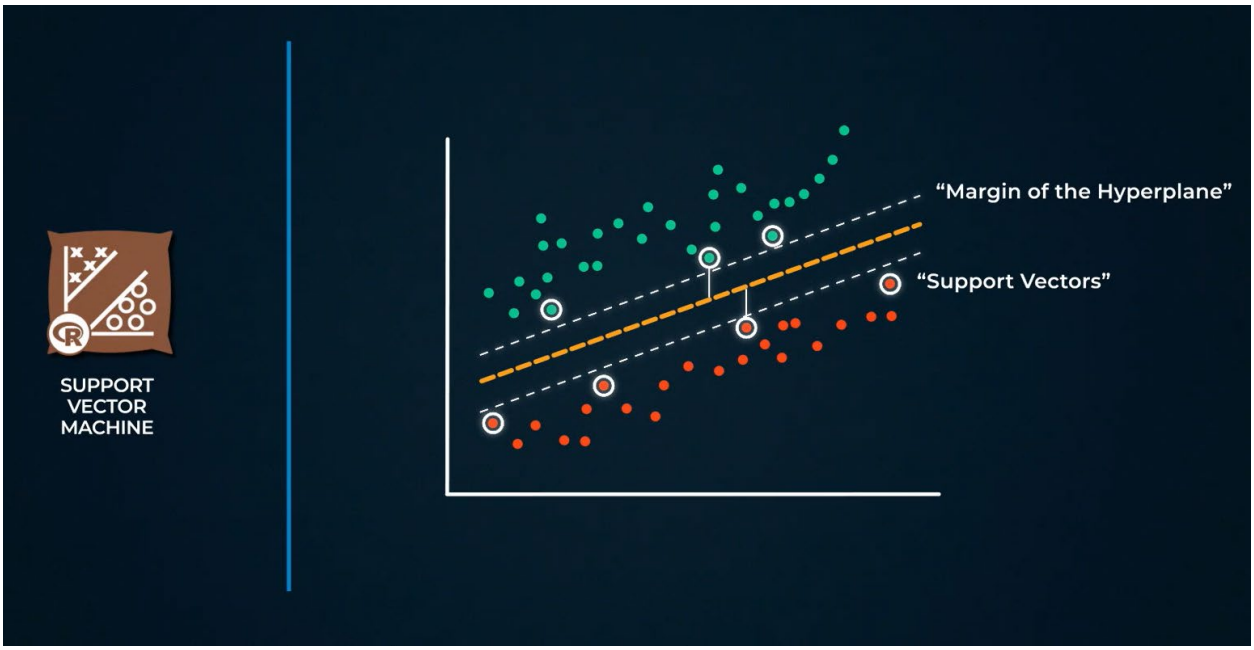


50. Support Vector Machine




SUPPORT
VECTOR
MACHINE
PROS
- Resistant to overfitting
- Can outperform other
algorithms on difficult classifications
CONS
- Computationally expensive
- Require fine-tuning


SUPPORT
VECTOR
MACHINE
"Hyperplane"
Instead of "line"
because it is usually more
than two dimensional



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

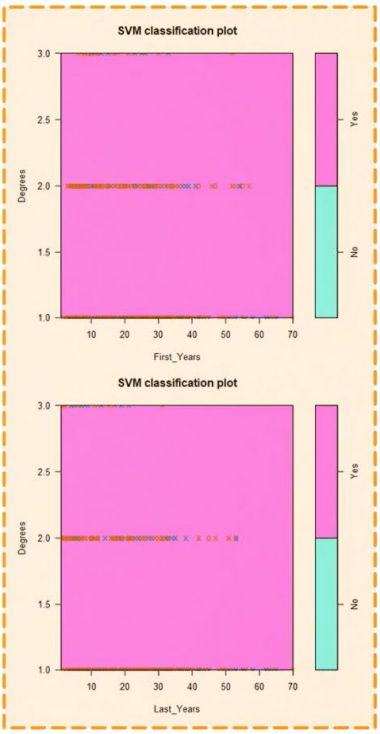
Confusion Matrix

	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:



51. Neural Network

PROS

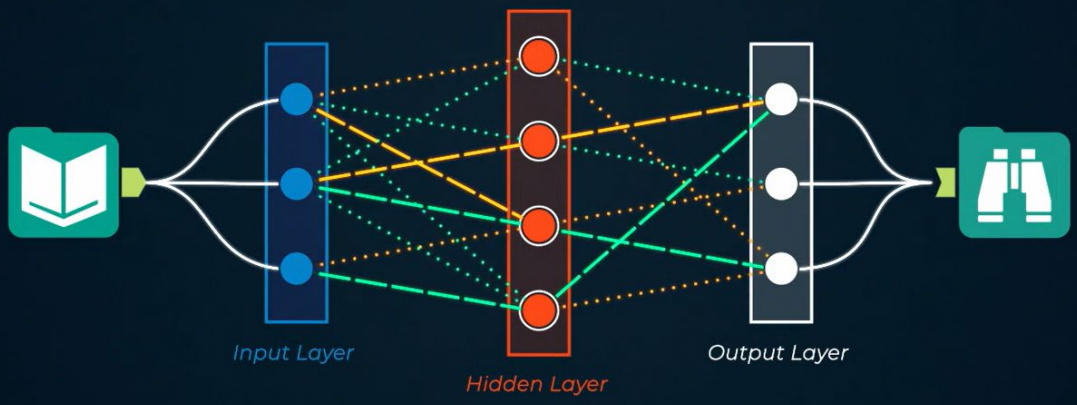
- Perform well for highly non-linear relationships
- Provides results quickly
- Resistant to multicollinearity

NEURAL NETWORK

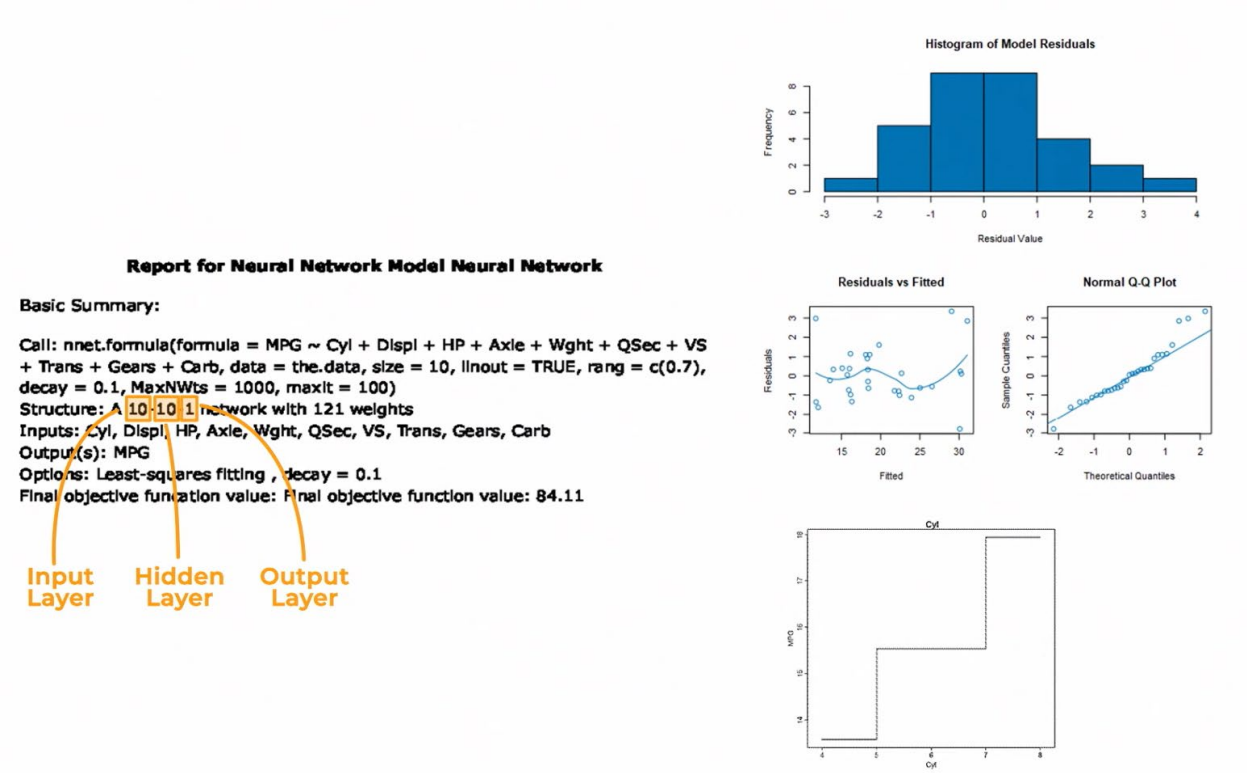
CONS

- Difficult to explain
- Sensitive to missing values
- Sensitive to outliers
- Time consuming to create

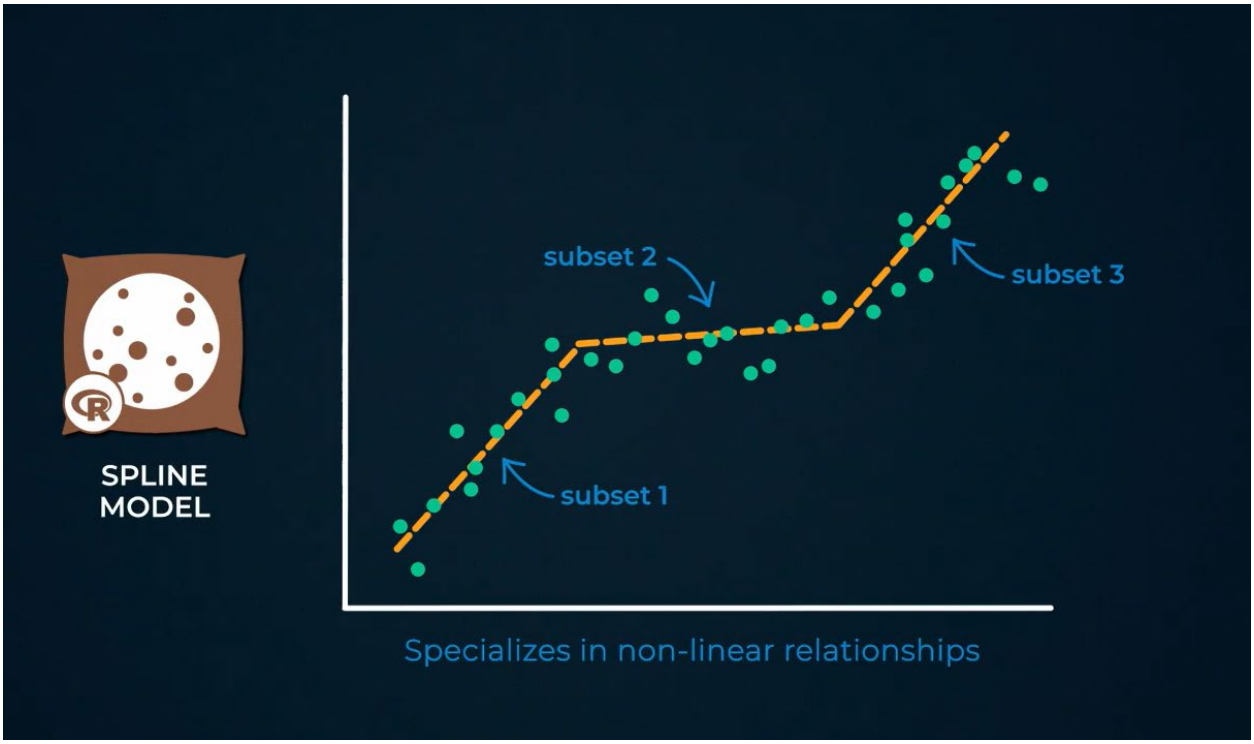
NEURAL NETWORK



You can specify the number of neurons
but connections, weights, and thresholds are automatic.



52. Spline model



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

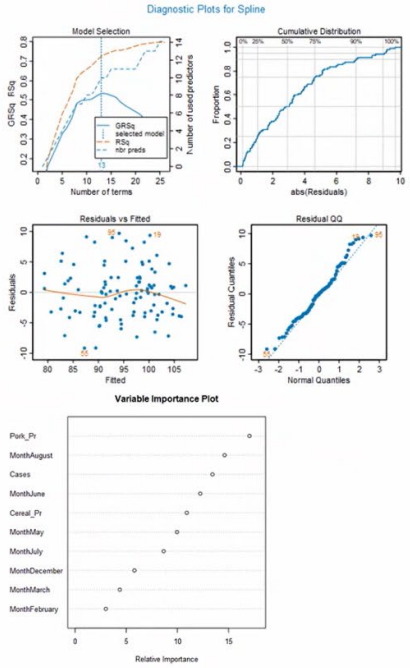
Selected 13 of 26 terms, and 10 of 20 predictors

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
(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

Continuous Predictors

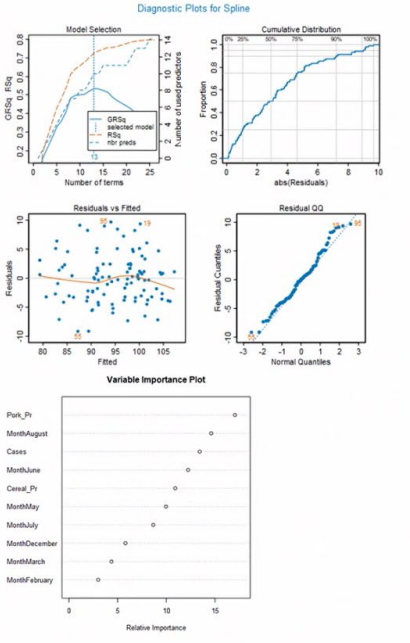
Selected 13 of 26 terms, and 10 of 20 predictors

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

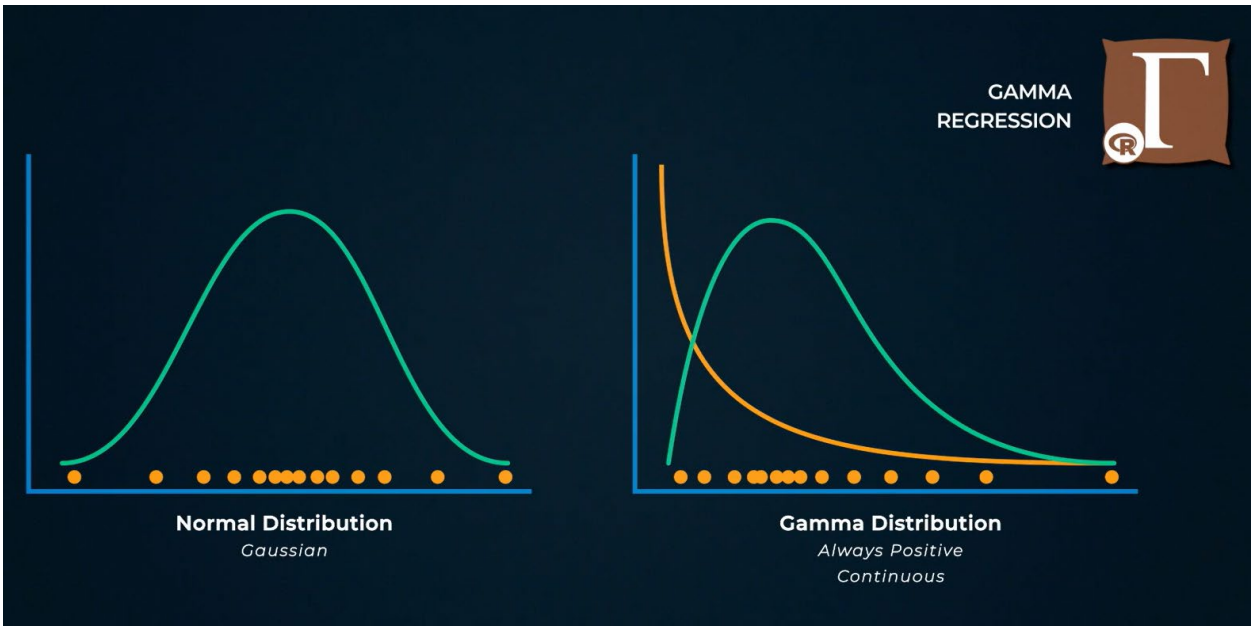
GCV 30.29 RSS 1818 GRSq 0.5323 RSq 0.7249

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

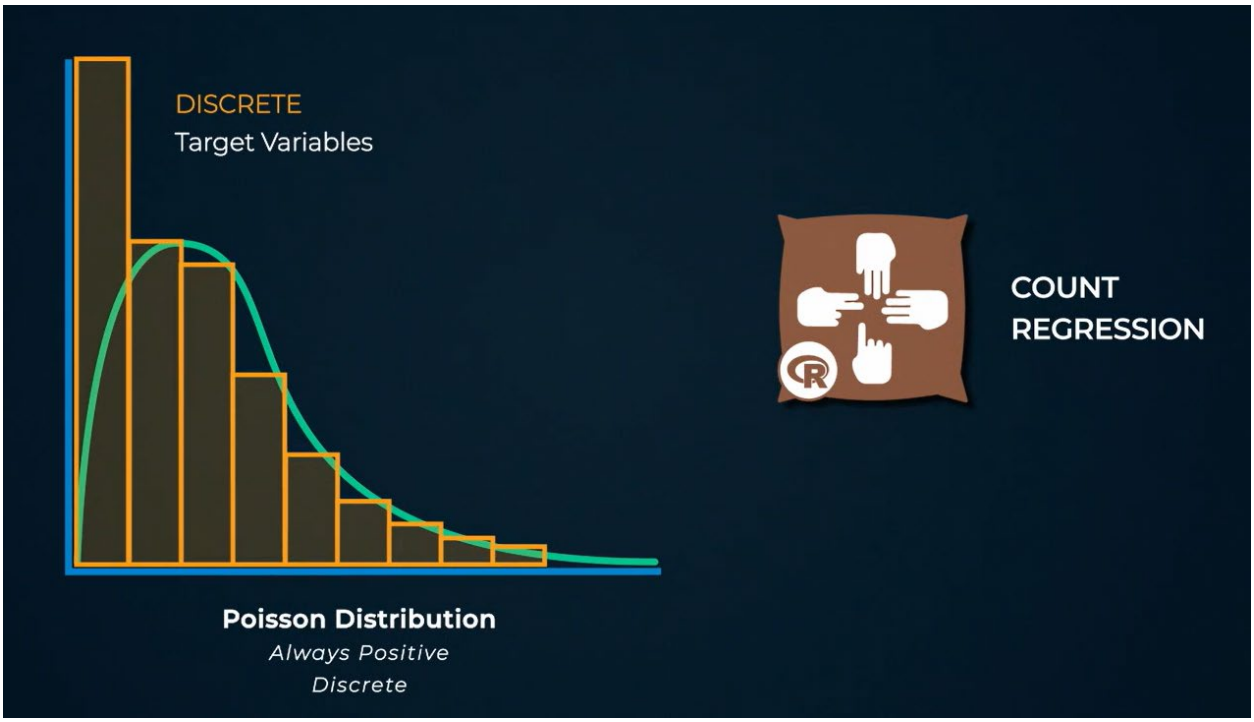
GLM Model McFadden R-Squared: 0.7249



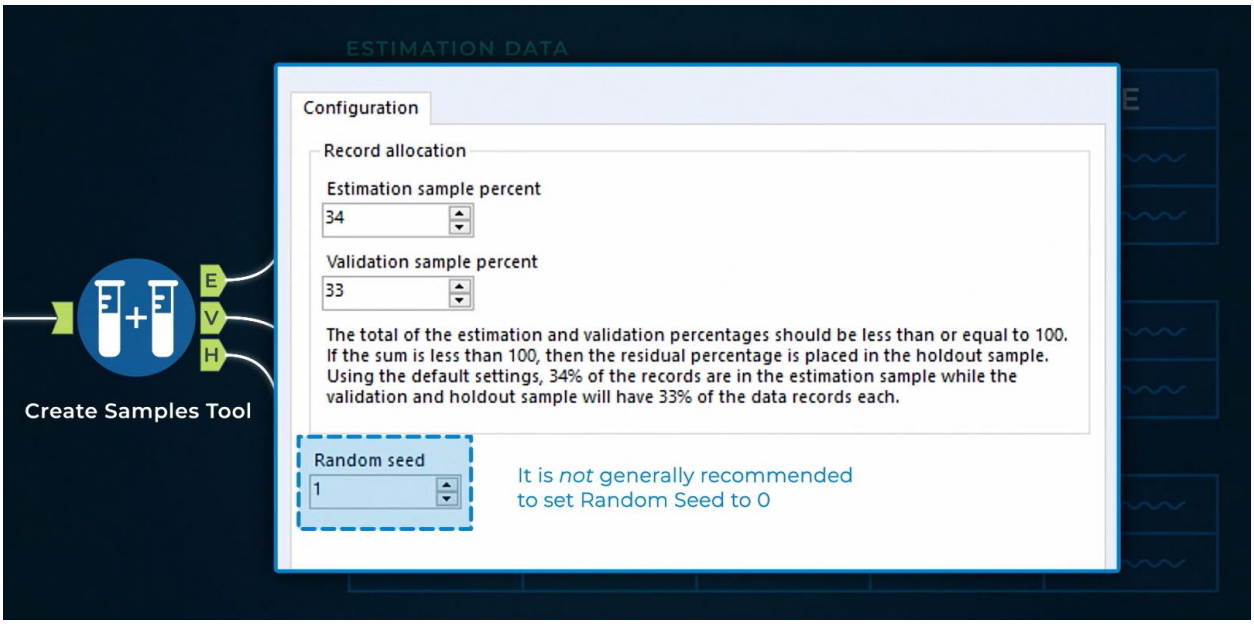
53. Gamma regression



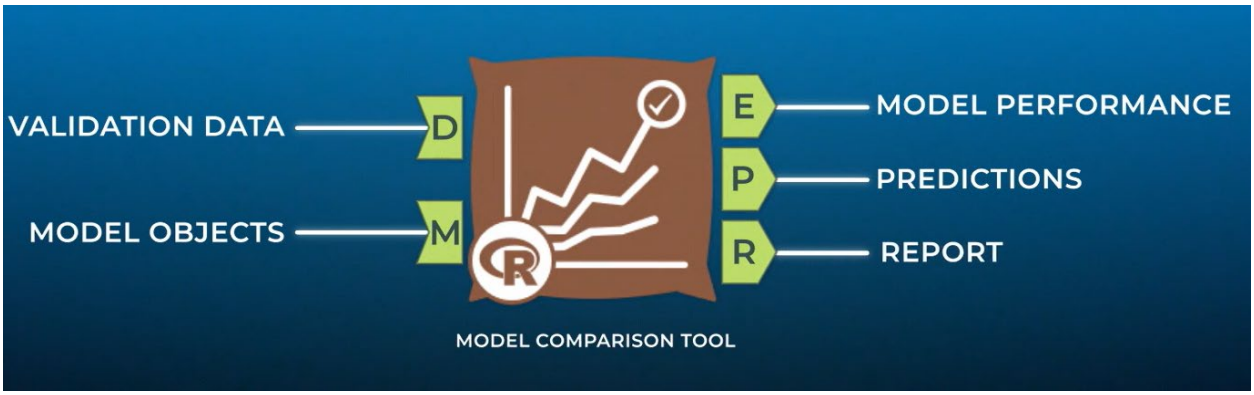
54. Count Regression

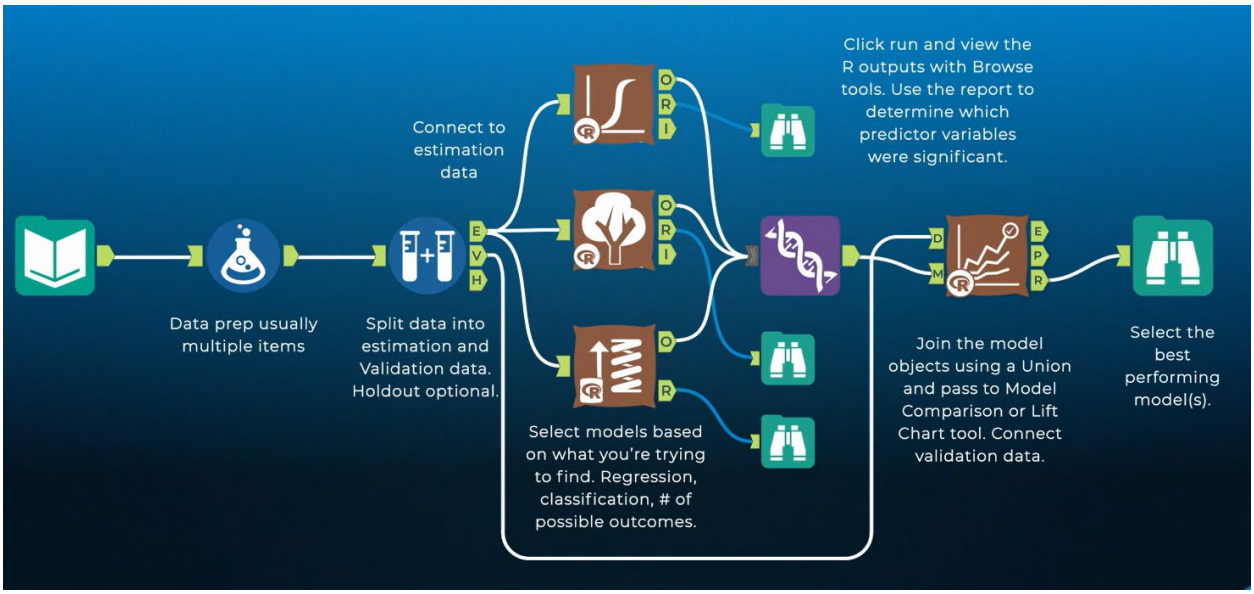


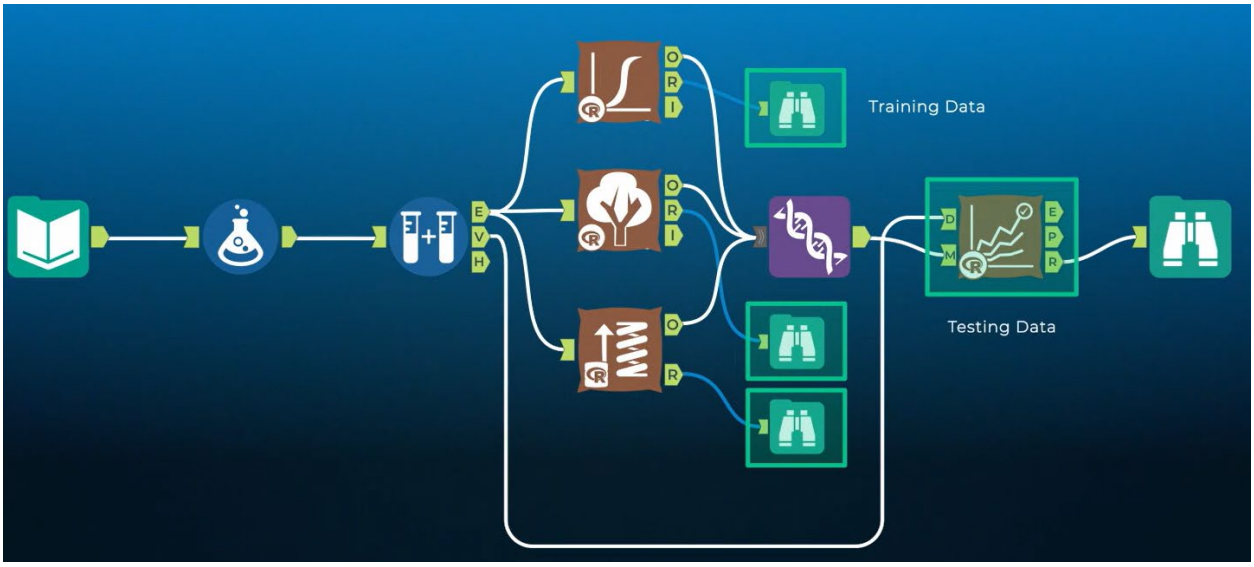
55. Create Samples. If you use a Random seed of `0`, then you make it impossible to generate the same results since randomly selected records will be output on each run of the workflow.

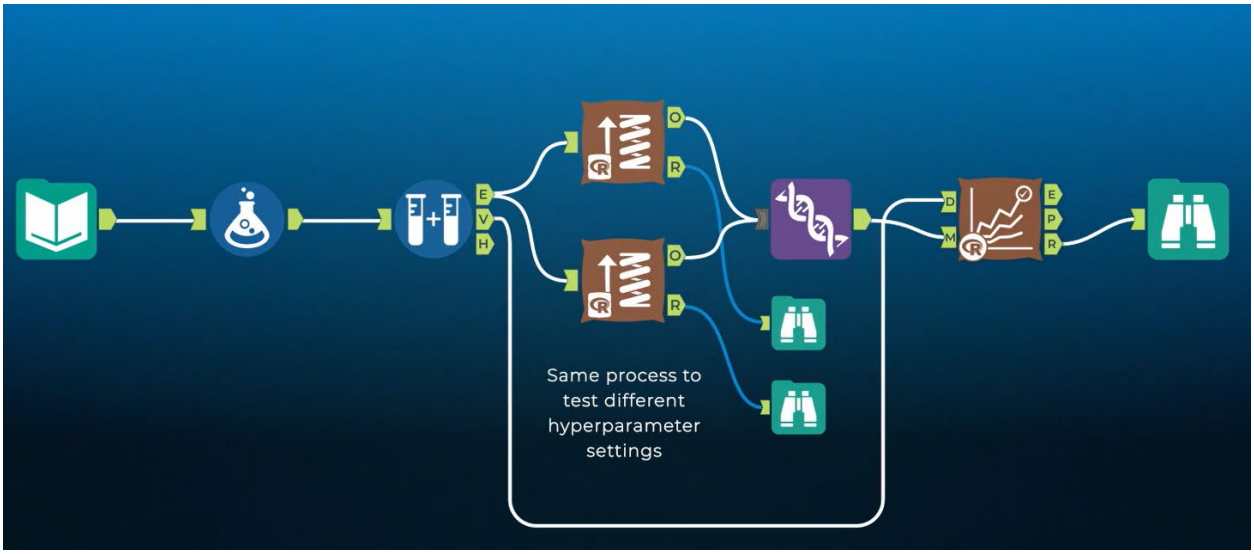


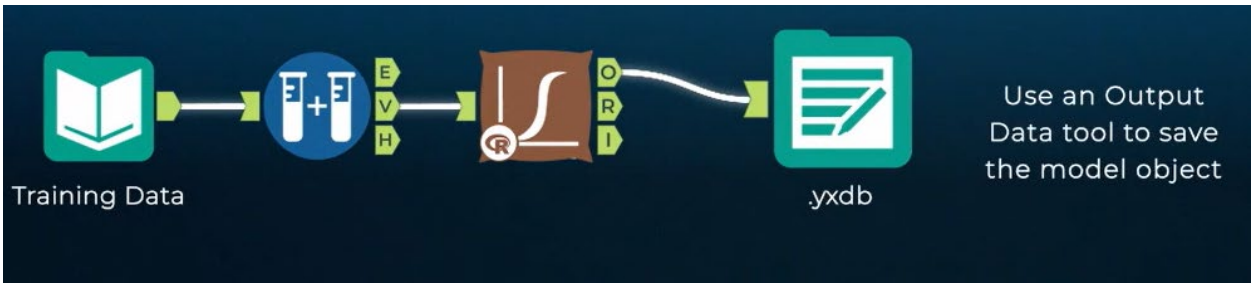
56. Model Comparison and Score

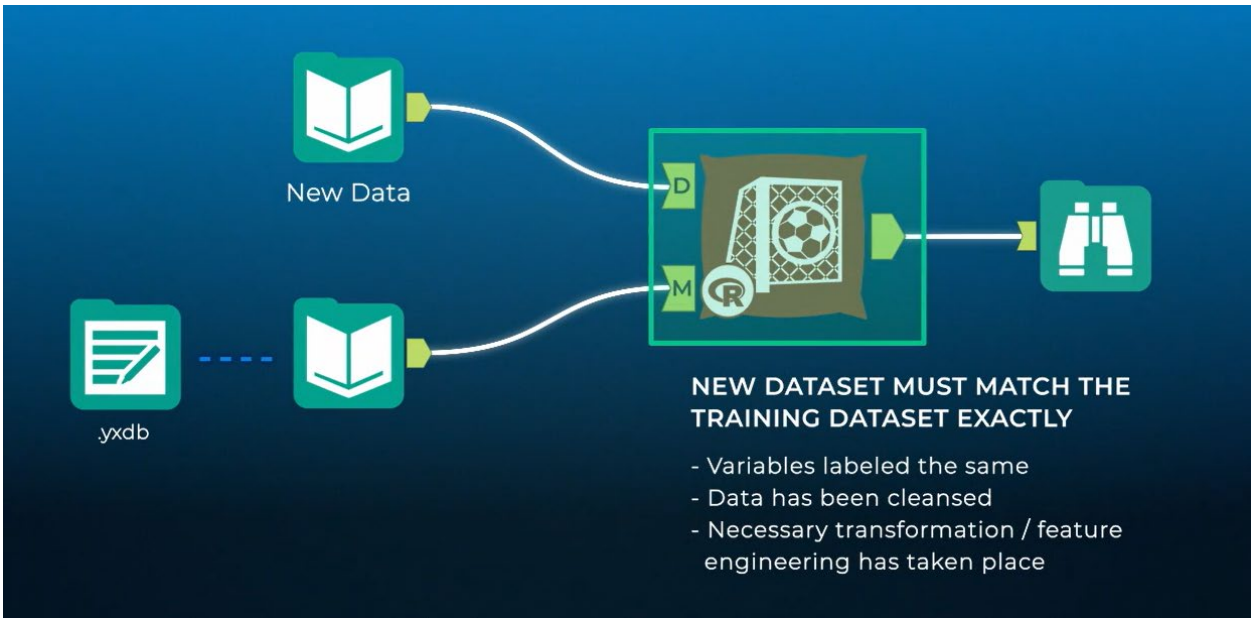




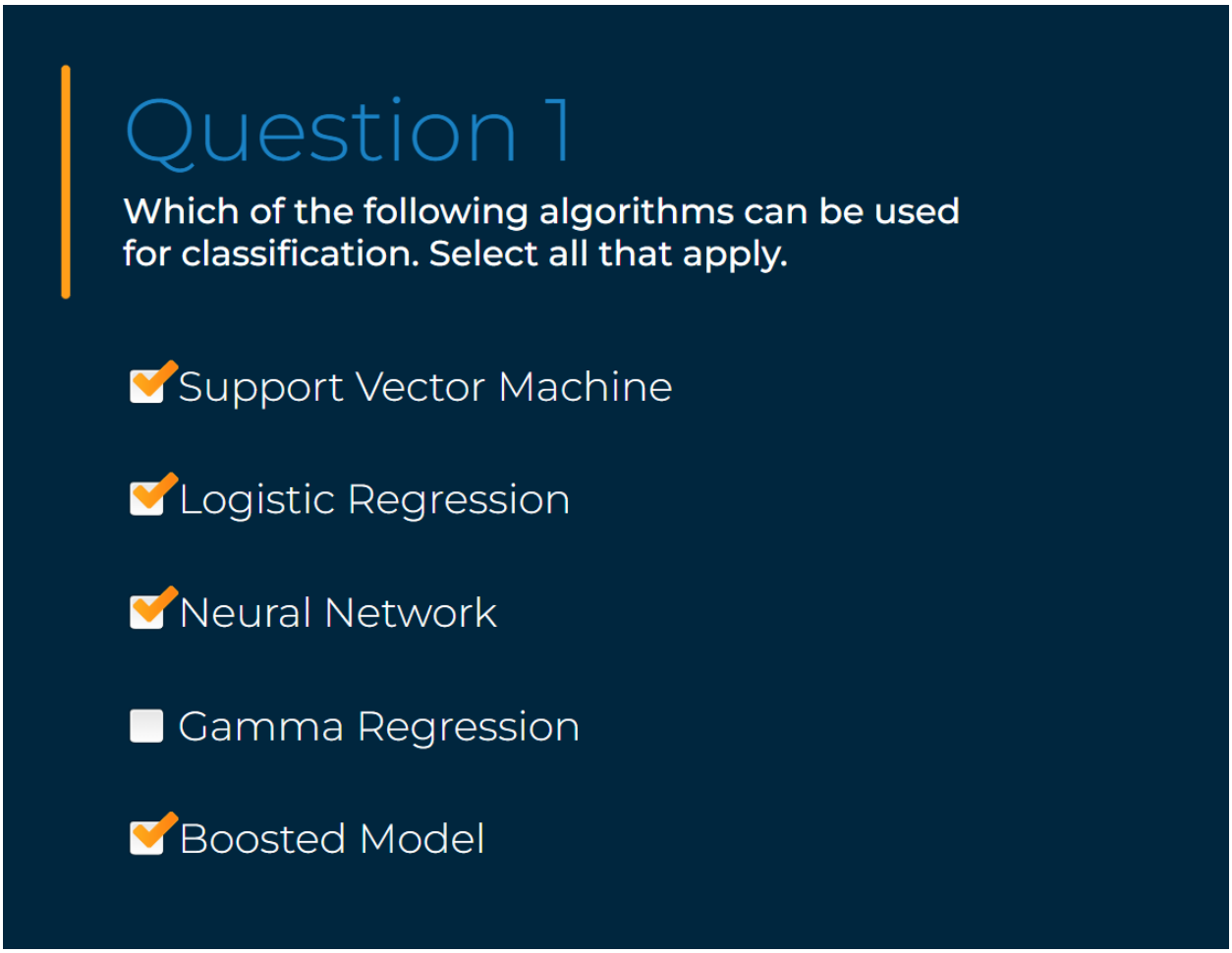

Training Data
Testing Data


Same process to test different hyperparameter settings


Training Data
.yxdb
Use an Output Data tool to save the model object



57. Predictive Questions



Question 2

Why would you split data into estimation and validation data?

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

- 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
- Depends on the dataset
- Support Vector Machine
- 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

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.

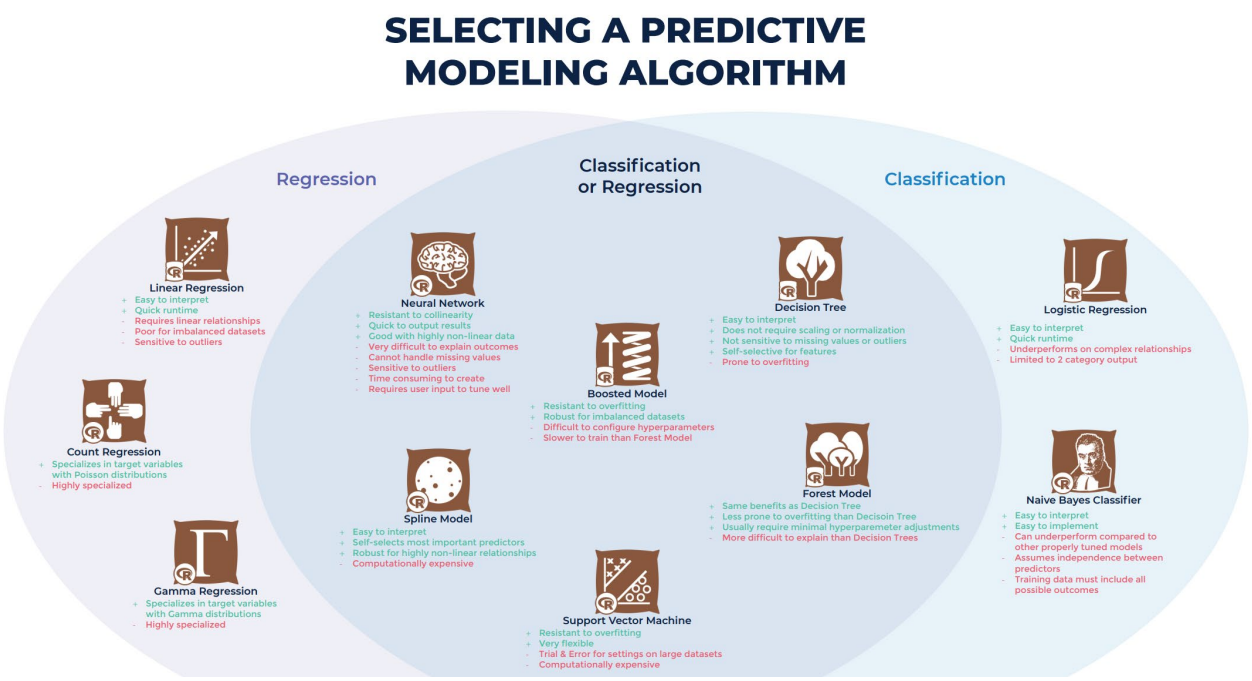
- [x] Your ability to explain the results of the model
- [x] The number of possible classifications
- [x] The datatype of the target variable
- [x] The distribution of the target variable

Question 10

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

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

58. Selecting a predictive modeling algorithm



59. Principal Component Analysis (PCA)

PRINCIPAL COMPONENT ANALYSIS

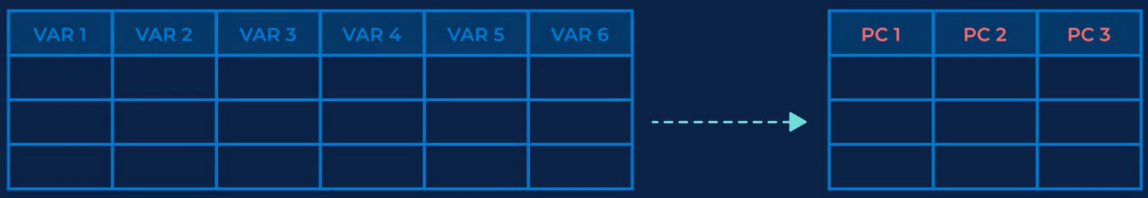
BENEFITS:

- Reduces multicollinearity
- Reduce dataset noise
- Determine the most important variables

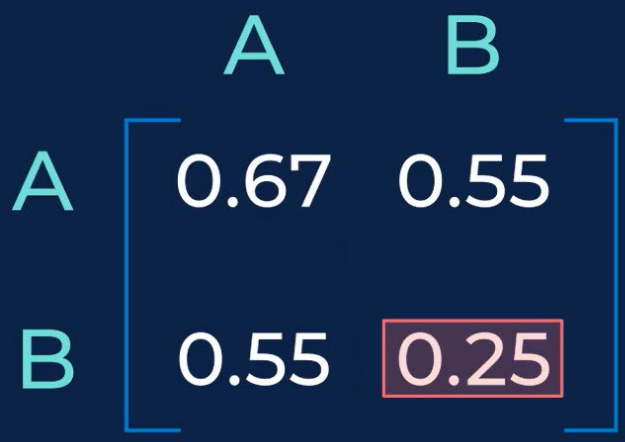
PCA can extract information from columns that have little utility on their own.

Variance does not *necessarily* indicate usefulness unless it is correlated with the target variable.

STEPS IN PCA



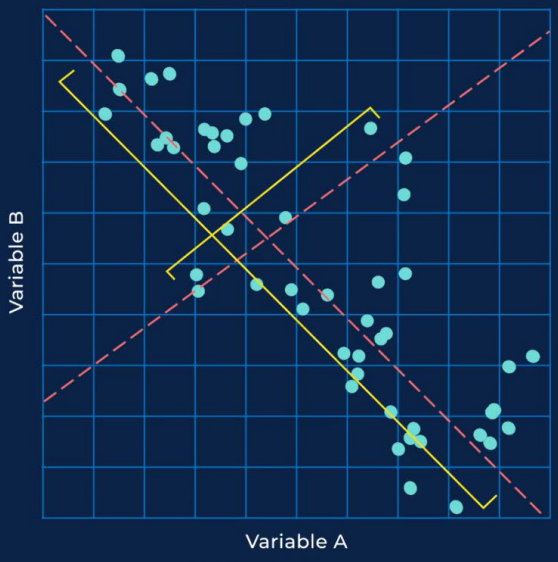
New variables are created from weighted combinations of existing variables.



COVARIANCE MATRIX

Describes how related two variables are

Not the same as a correlation matrix



EIGENDECOMPOSITION

$$\begin{bmatrix} -1 & 4 \\ 0 & 5 \end{bmatrix} * \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 \\ 15 \end{bmatrix} = 5 * \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

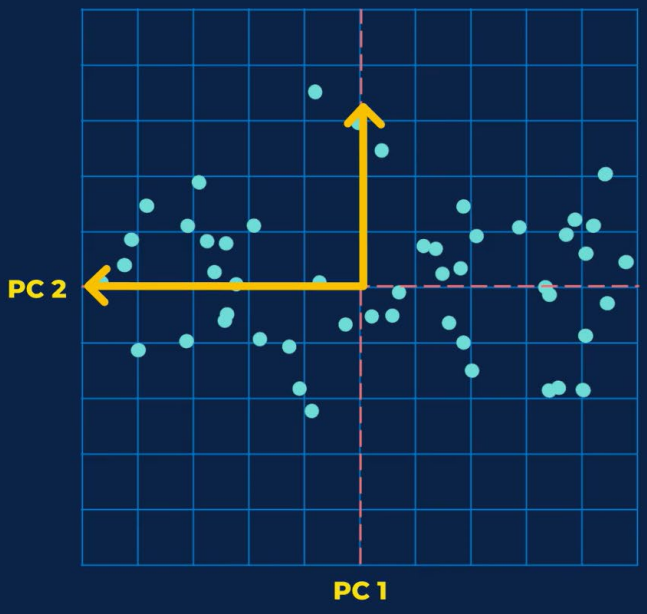
Eigenvalue — Eigenvector

Ranked eigenvalues are used to select the most important variables.

Eigenvectors are used to transform the dataset to reflect new principal components.

Rather than predicting a target variable, PCA determines a NEW independent variable.

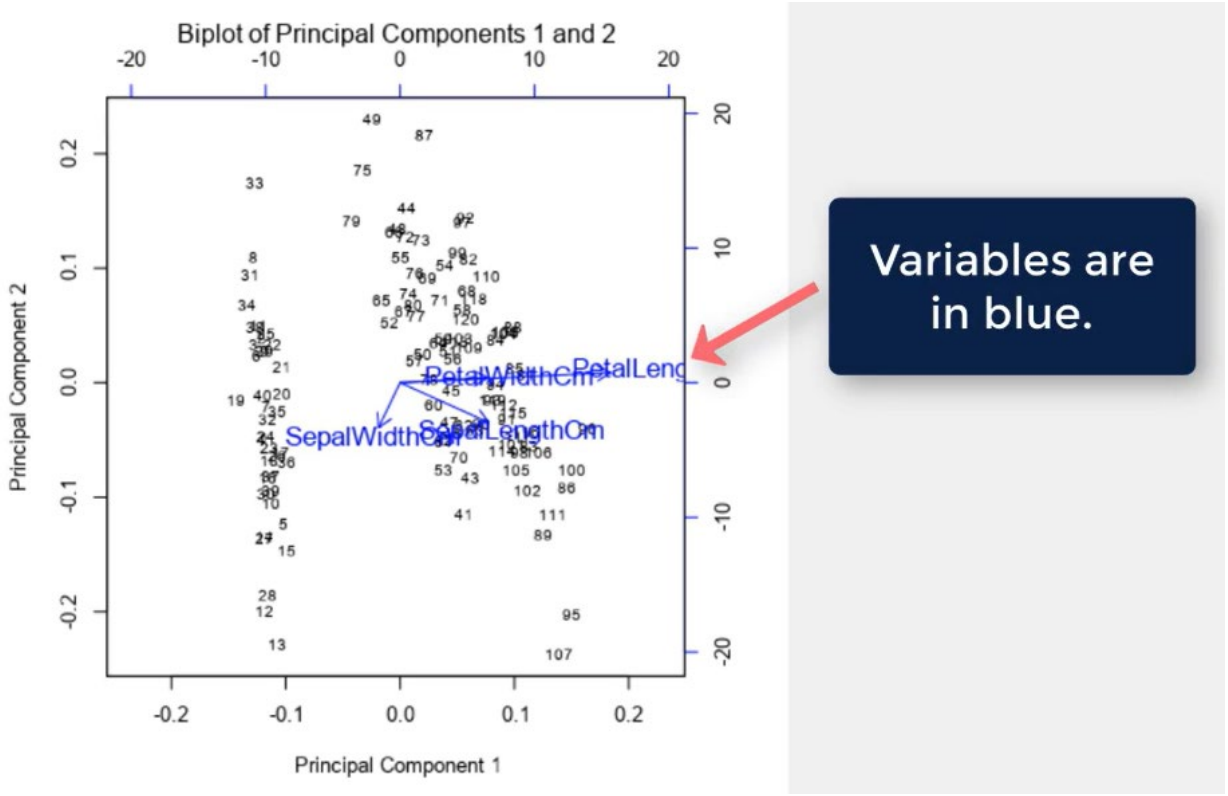
PROJECTION TO NEW COMPONENTS



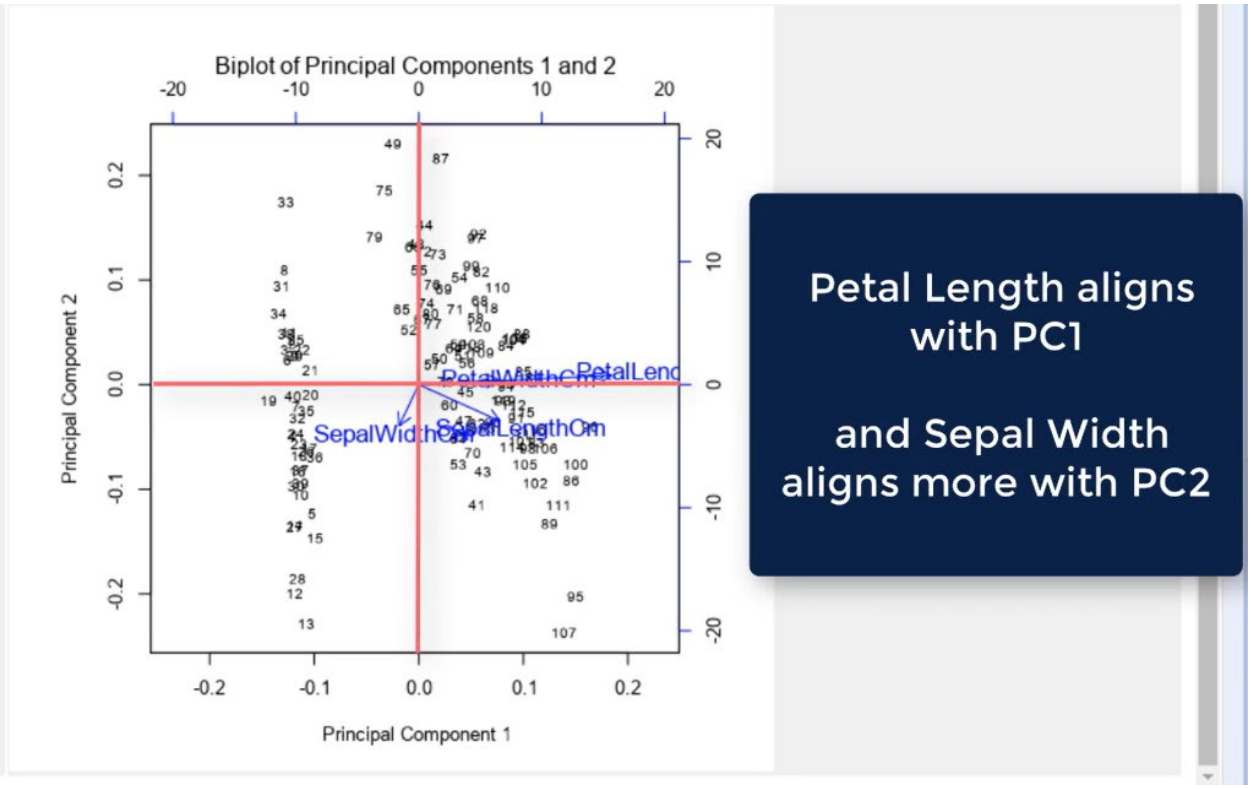
TRANSFORMATION

All principal components are described as orthogonal

New variables (PCs) are entirely independent of each other



Variables that align with an axis have a strong impact on that Principal Component.



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.

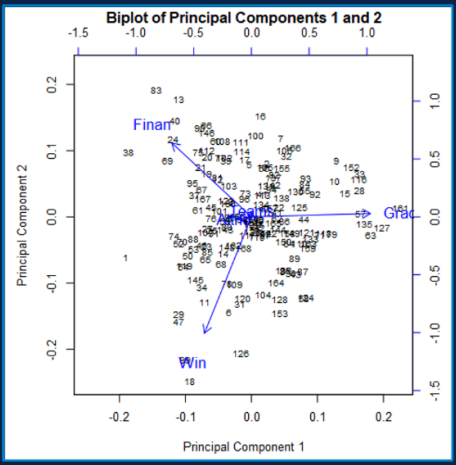
QUESTION 1

PCA performs dimensional reduction by:

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

QUESTION 2

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



- Grad
- Finan
- Not enough info provided
- Win

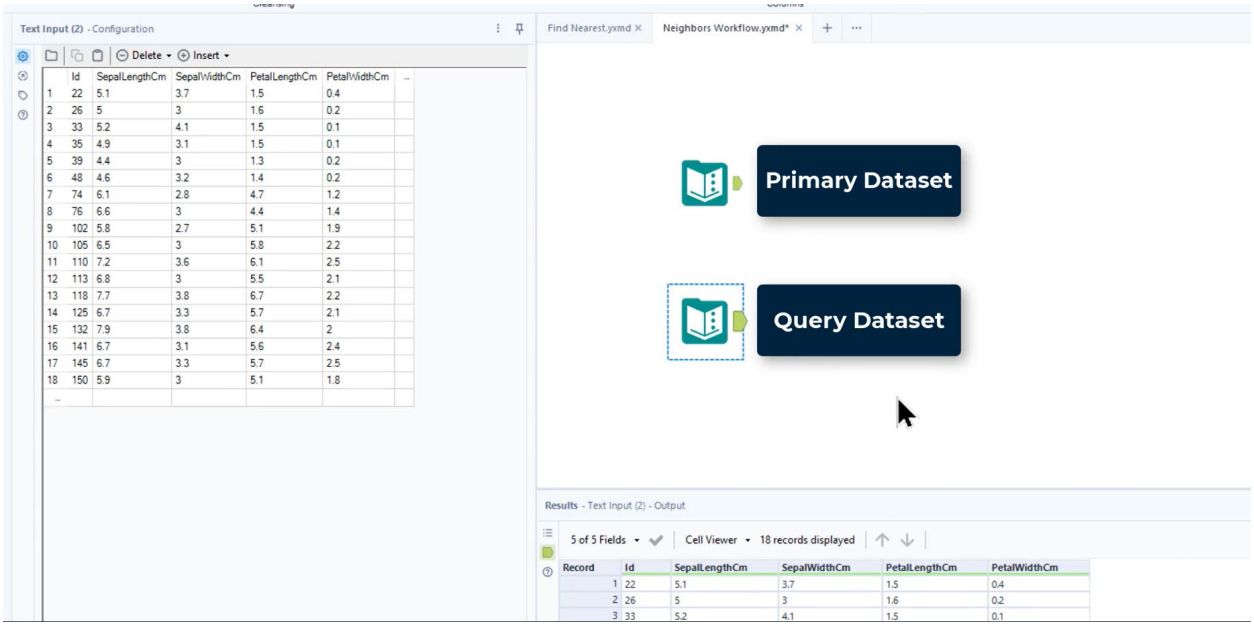
QUESTION 3

Which of the following are potential issues when using PCA? (select all that apply)

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

60. Find Nearest Neighbor


FIND NEAREST NEIGHBORS
Returns a list of closest data points
Calculates Euclidean distance based on associated values
5.39
2
5
FIND NEAREST
Finds physical distance between spatial objects


Text Input (2) - Configuration
Delete
Insert
Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
1 22 5.1 3.7 1.5 0.4
2 26 5 3 1.6 0.2
3 33 5.2 4.1 1.5 0.1
4 35 4.9 3.1 1.5 0.1
5 39 4.4 3 1.3 0.2
6 48 4.6 3.2 1.4 0.2
7 74 6.1 2.8 4.7 1.2
8 76 6.6 3 4.4 1.4
9 102 5.8 2.7 5.1 1.9
10 105 6.5 3 5.8 2.2
11 110 7.2 3.6 6.1 2.5
12 113 6.8 3 5.5 2.1
13 118 7.7 3.8 6.7 2.2
14 125 6.7 3.3 5.7 2.1
15 132 7.9 3.8 6.4 2
16 141 6.7 3.1 5.6 2.4
17 145 6.7 3.3 5.7 2.5
18 150 5.9 3 5.1 1.8
Find Nearest.yxmd
Neighbors Workflow.yxmd*
Primary Dataset
Query Dataset
Results - Text Input (2) - Output
5 of 5 Fields
Cell Viewer
18 records displayed
Record Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
1 22 5.1 3.7 1.5 0.4
2 26 5 3 1.6 0.2
3 33 5.2 4.1 1.5 0.1

...find Neighbors from this dataset

For datapoints in this dataset...

Datasets should have some fields in common:

One key field to identify datapoints

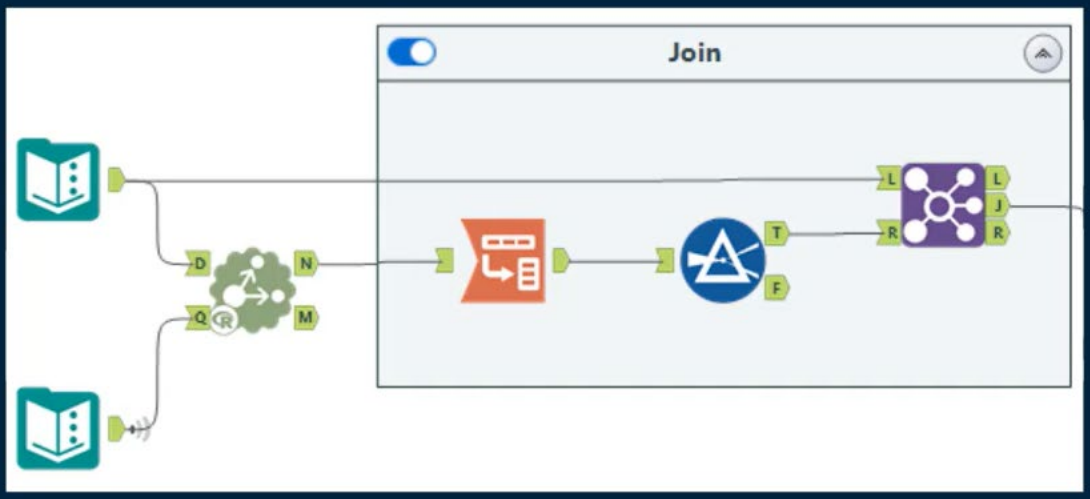
Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
22	5.1	3.7	1.5	0.4
26	5	3	1.6	0.2
33	5.2	4.1	1.5	0.1
35	4.9	3.1	1.5	0.1
39	4.4	3	1.3	0.2
48	4.6	3.2	1.4	0.2
74	6.1	2.8	4.7	1.2
76	6.6	3	4.4	1.4
102	5.8	2.7	5.1	1.9

and numeric fields that share naming conventions.

Non-numeric fields are dropped.

For a List of Neighbors:



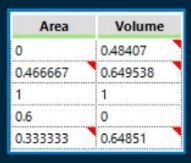
Results - Join (6) - Out - Join

3 of 3 Fields | Cell Viewer | 90 records displayed

Record	Querry ID	Neighbor ID	Species
1	22	1	Iris-setosa
2	26	2	Iris-setosa
3	35	2	Iris-setosa
4	39	3	Iris-setosa
5	48	3	Iris-setosa
6	22	5	Iris-setosa
7	39	7	Iris-setosa
8	48	7	Iris-setosa
9	39	9	Iris-setosa

Question 1

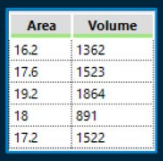
Given the configuration options below, which of these options would you expect to see from the M output anchor?

- 

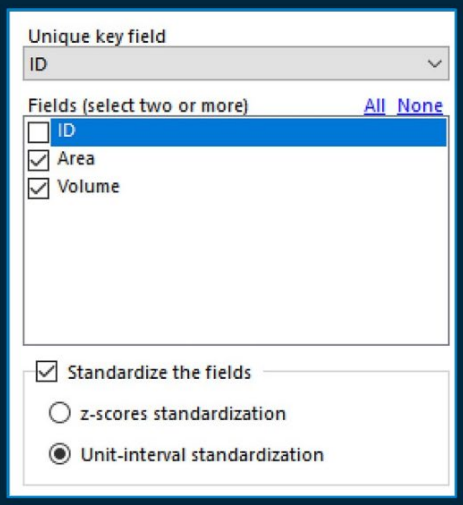
Area	Volume
0	0.48407
0.466667	0.649538
1	1
0.6	0
0.333333	0.64851

- 

Area	Volume
-1.310174	-0.199005
-0.036394	0.256106
1.419355	1.220035
0.327544	-1.530415
-0.400331	0.253279

- 

Area	Volume
16.2	1362
17.6	1523
19.2	1864
18	891
17.2	1522



Question 2

Which of these algorithms calculates the exact distance between all datapoints in the datasets?

- Cover Tree
- KD-Tree
- VR
- CR
- Linear Search

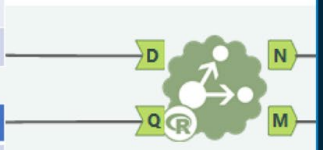
Question 3

Given these two datasets, will the tool error?

- Yes
- No

ID	Area	Volume	Category
1	17.2	1234	A
2	16.4	1532	A
3	32.1	3546	C

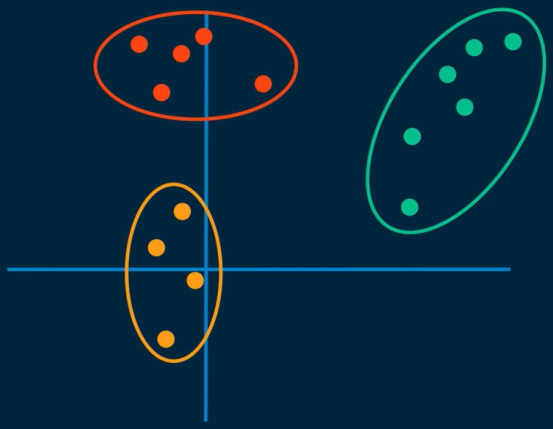
ID	Area	Volume
5	15.2	1730
6	21.8	1695



61. Clusters

CLUSTERS

Groupings of similar datapoints and separation of dissimilar datapoints.



STEPS

User inputs K = 2

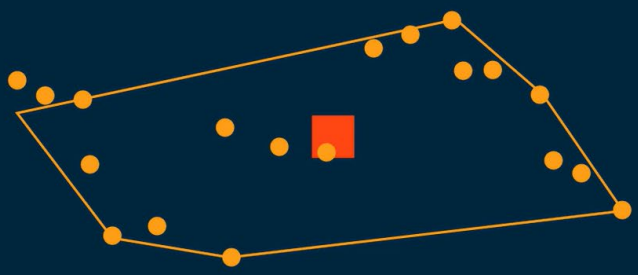
2 randomly selected points are assigned as centroids.

The distance of all other points are calculcated to the centroids, then assigned to the closest centroid.

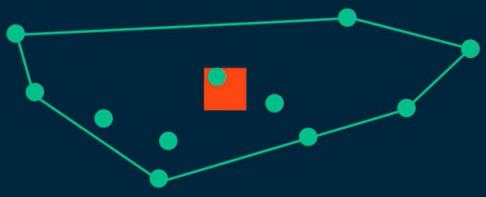
Once all points have been assigned, the centroids are re-calculated.

The process of classifying each point is repeated with calculations based on the new centroids.

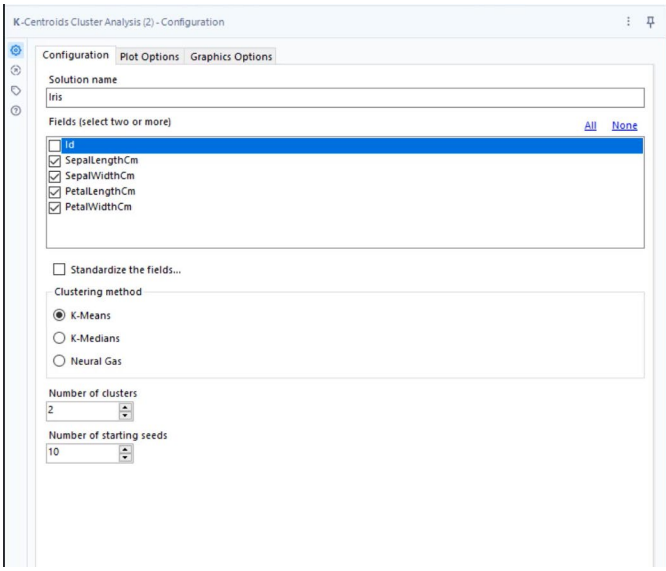
DATA



"CONVERGENCE"



For optimal results, the algorithm runs several times with randomly selected starting points.



Select Variables

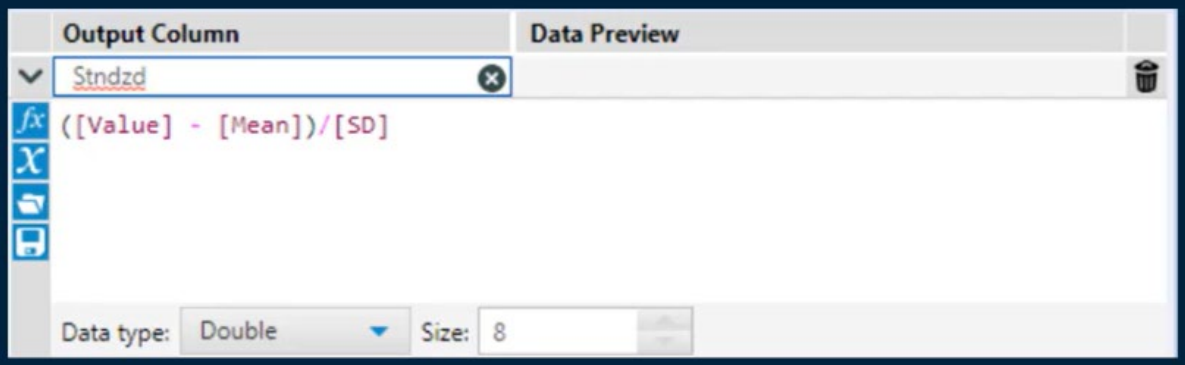
Only numeric columns are displayed

Minimum of 2 variables selected

Cannot contain Null values

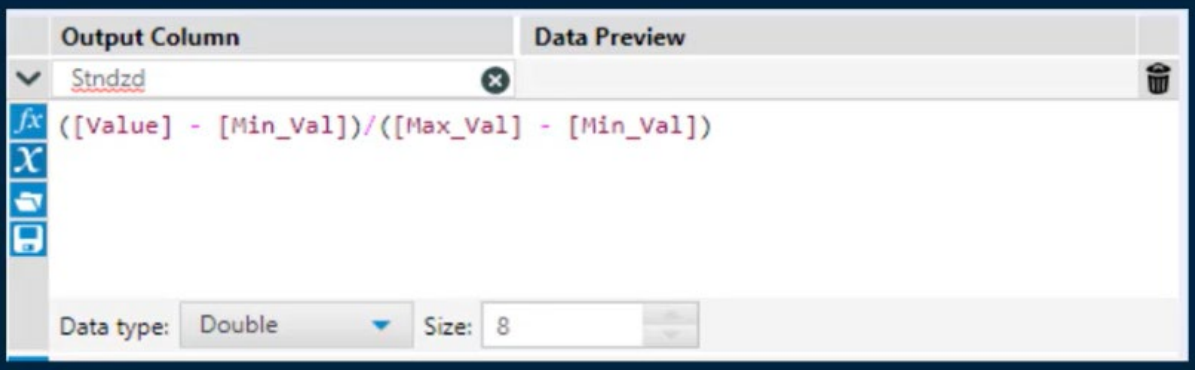
Sensitive to outliers

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

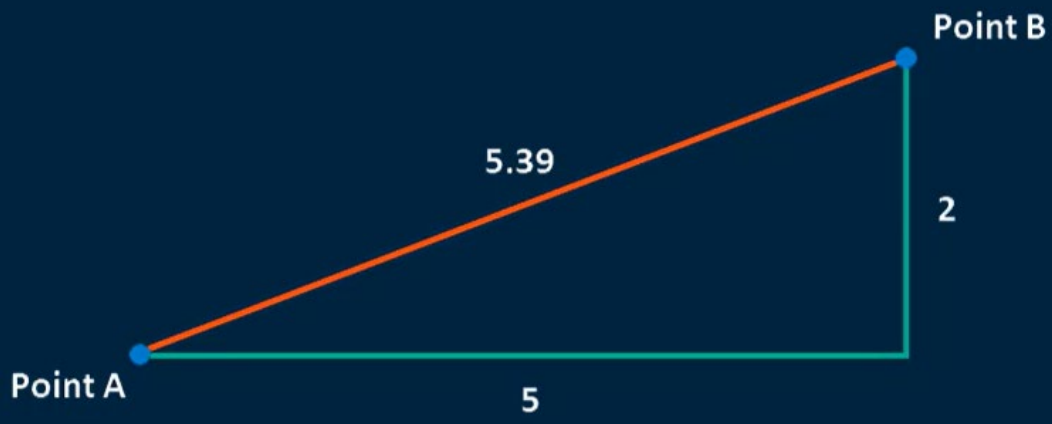
K-Means

Euclidean distance

Centroids from mean values

K-Medians

Manhattan distance = 7



K-Medians

Manhattan distance

Centroids from median values

Neural Gas

Euclidean distance

Centroids from weighting of all values

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.

Append Cluster

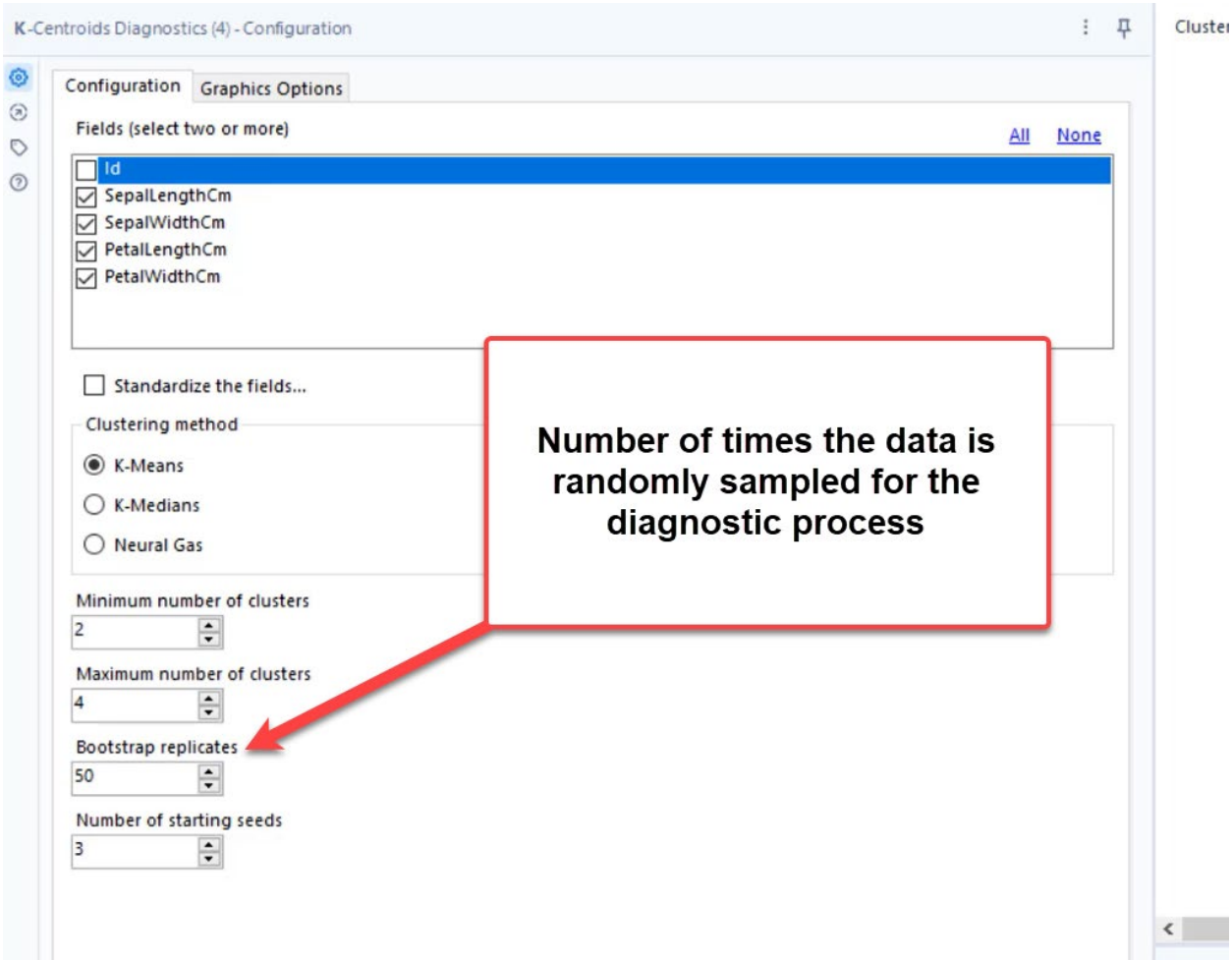
Attach the model object to the Append Cluster tool to "score" datasets.

CONSIDERATIONS

- Once configured, K-Centroids Cluster Analysis will always return results.
  Be mindful of your inputs.

- All configuration options and provided datapoints impact clustering results.
  Be consistent wherever possible.

- It is difficult to know the best settings for your dataset.
  e.g. the number of clusters.



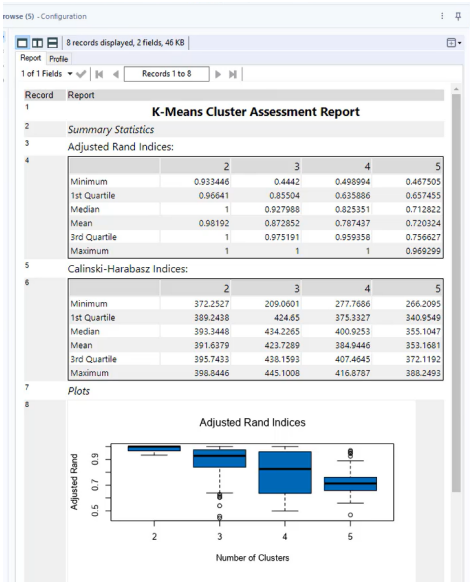
Bootstrap Replication

Randomly selects and duplicates data values.

A model is created for each of the number of clusters in the range on each subset of data.

This provides a better idea of the dataset's characteristics to improve accuracy.

It also increases runtimes.



Adjusted Rand Index

Similarity of models with equal number of clusters

-across bootstrap replicates

Range from 0-1

-higher values are desirable

"Consistency between runs"

Calinski-Harabasz Index

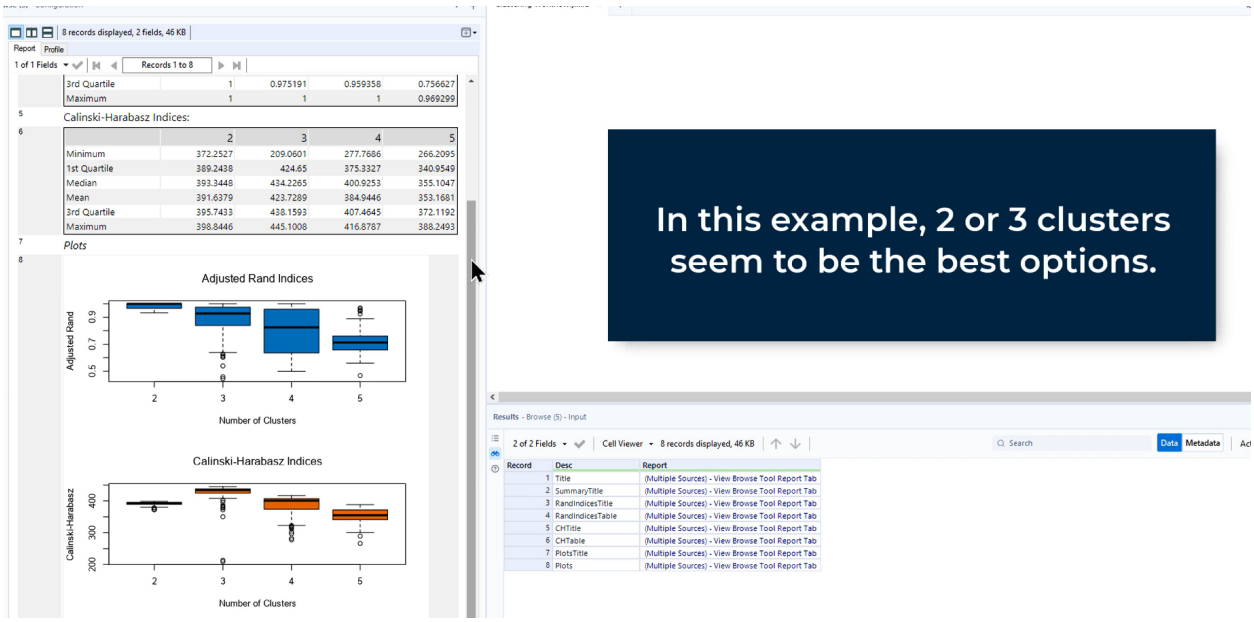
Definition of the clusters

-similarity of like datapoints vs
differences between clusters

Ratio:

-numerator = separation of clusters
-denominator = tightness in clusters

Higher values are desirable



Question 1

What does the number of starting seeds configuration control?

- The number of clusters that will be created
- The number of times the algorithm is run to convergence
- The number of columns used to calculate centroids
- The number of records used in creating the model

Question 2

What is the difference between K-Means and K-Medians?
Select all that apply.

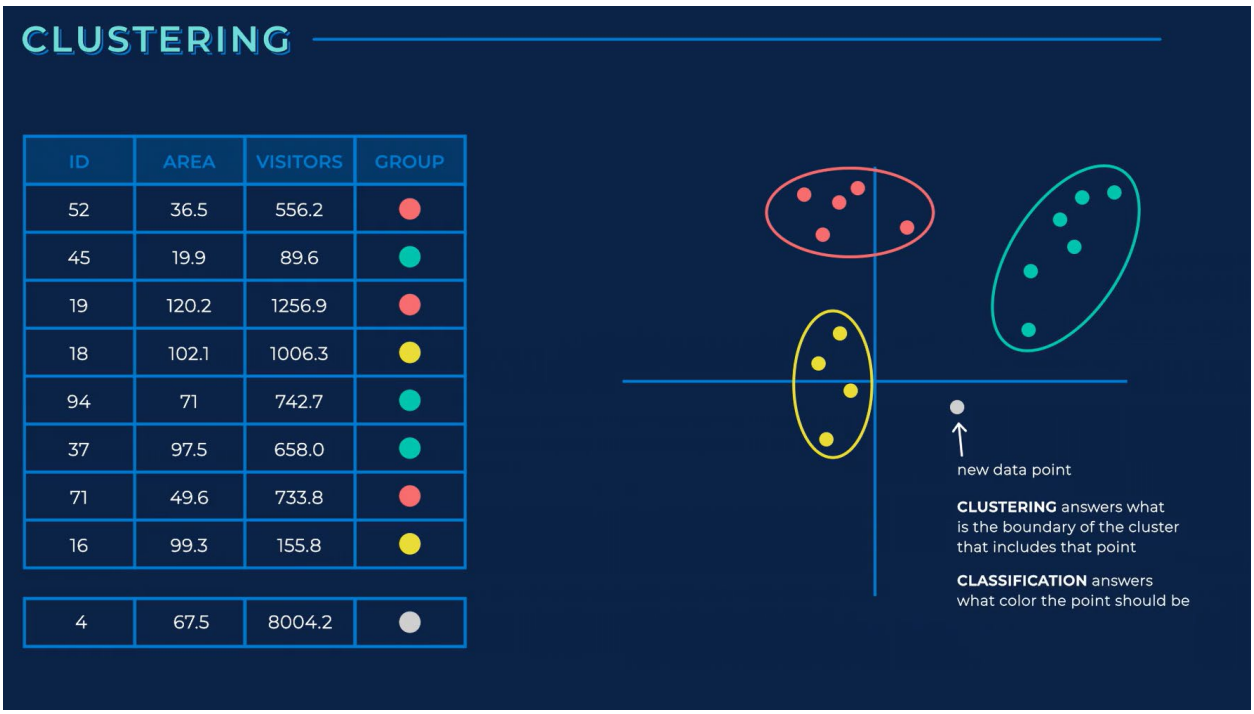
- [x] K-Means uses mean values to calculate centroids
- [x] K-Medians uses median values to calculate centroids
- [x] K-Medians uses Manhattan distance and K-Means uses Euclidean distance
- [ ] K-Medians uses Euclidean distance and K-Means uses Manhattan distance

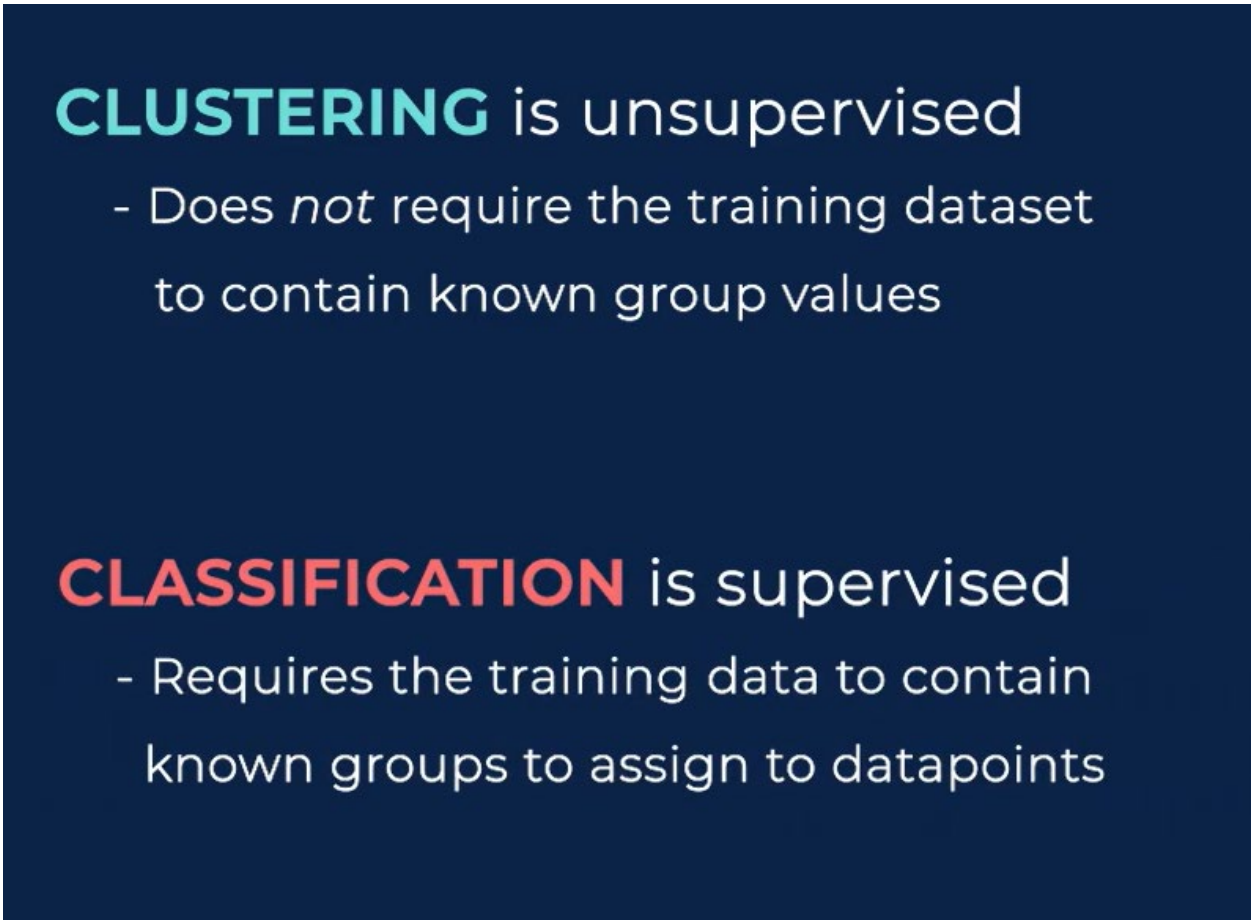
Question 3

When using the K-Centroids Diagnostics tool, what indicates a better performing model? Select all that apply.

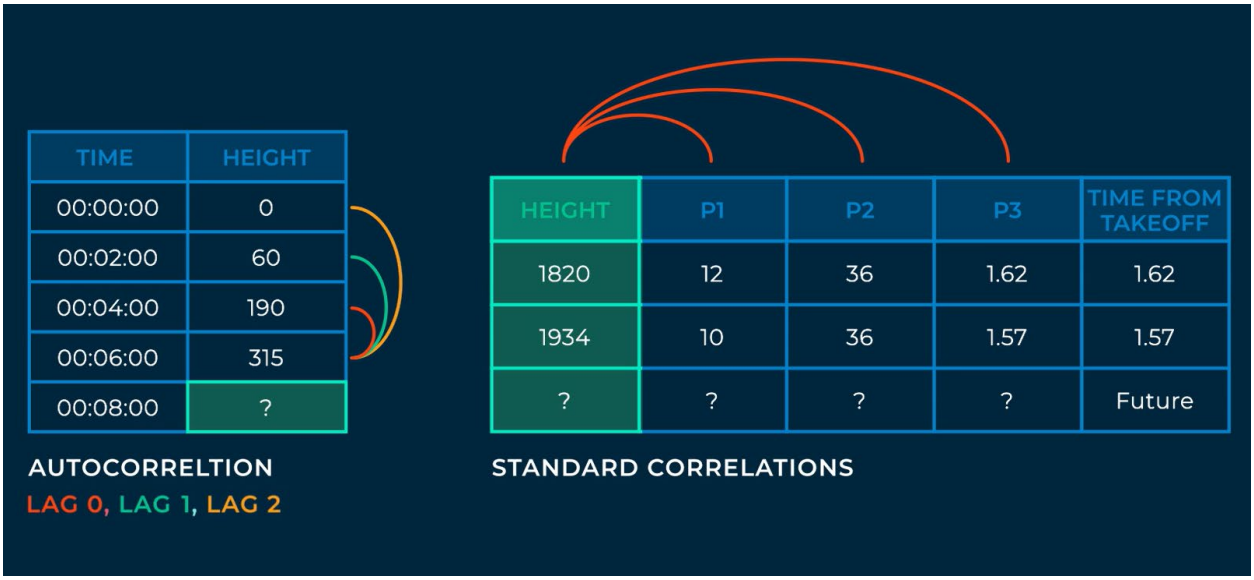
- [x] Compressed box plots with high means on the Calinski-Harabasz Index
- [ ] Larger box plots on either index
- [x] Compressed box plots with high means on the Adjusted Rand Index
- [ ] Compressed box plots with low means on the Adjusted Rand Index
- [ ] Compressed box plots with low means on the Calinski-Harabasz Index

62. Clustering vs Classification



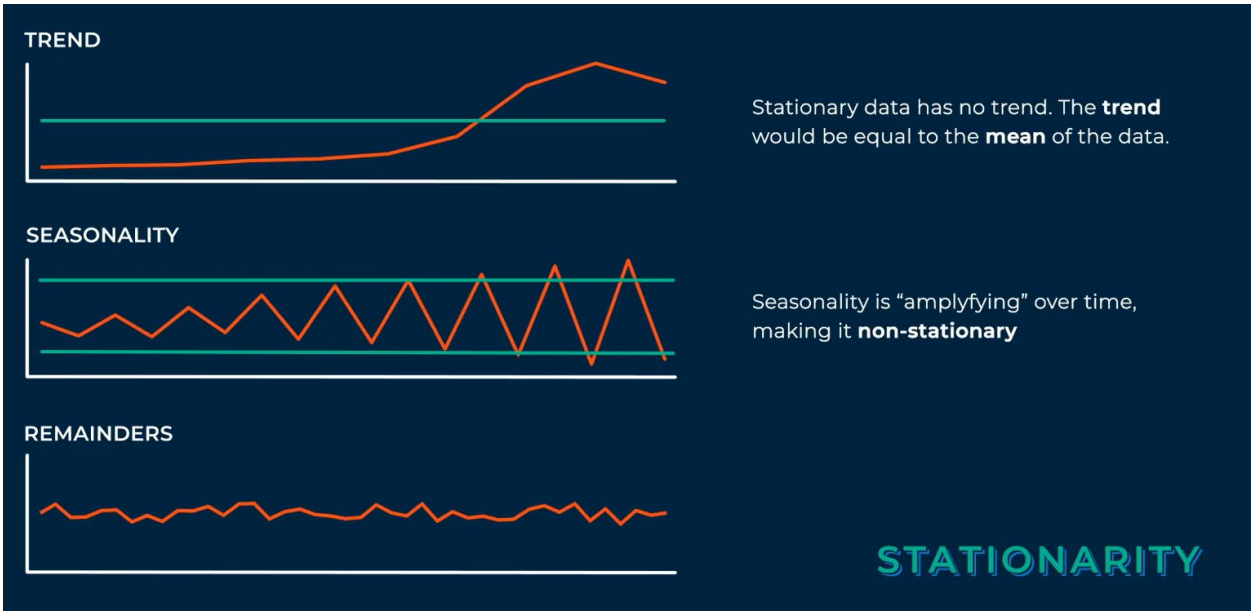


63. Time Series Forecasting



By examining the relationship of a variable to itself at various lags, we can find patterns which can be extrapolated into the future.


DECONSTRUCTION
WORKLOAD
MONDAY
TUESDAY
WEDNESDAY
THURSDAY
FRIDAY
MONDAY
TUESDAY
WEDNESDAY
THURSDAY
FRIDAY
TREND
direction
SEASONALITY
REMAINDERS
unexplained factors


TREND
SEASONALITY
REMAINDERS
Stationary data has no trend. The **trend** would be equal to the **mean** of the data.
Seasonality is “amplyfying” over time, making it **non-stationary**
STATIONARITY

REGULARLY SPACED INTERVALS

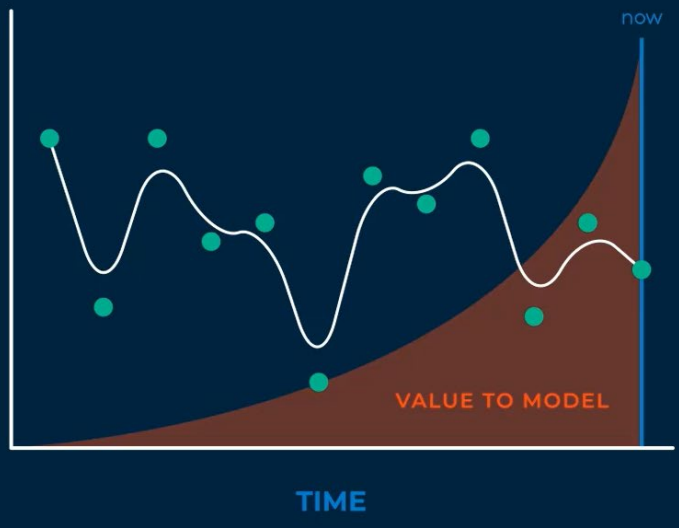
DATE	QUANTITY	UNIT
2014-05-10	1372	kg
2015-05-10	1406	kg
2016-05-10	1525	kg
2017-06-10	1634	kg
2018-05-10	1892	kg
2019-05-10	1750	lb
2020-05-10	2039	lb

Irregular Month

Inconsistent Units

Plotting these togeher **breaks** assumptions made by the algorithm

SOME DATA IS MORE VALUABLE



Recent data is usually the most valuable, so order is very important.

TIME SERIES MODELS	PREDICTIVE MODELS
Short life span	Potentially infinate life span
Projecting further into the future reduces certainty	Unclear relationships between variables reduces certainty
Model assuptions change with each new data point	The time when observations occur is irrelevant

Question 1

True or False? The primary purpose of Time Series is to forecast what will happen in the future, not discover which factors will contribute to that outcome.

- True
- False

Question 2

True or False? Time series forecasting is entirely dependent on historical values.

- True
- False

Question 3

True or False? Time Series models have relatively short lifespans when compared to other types of machine learning models.

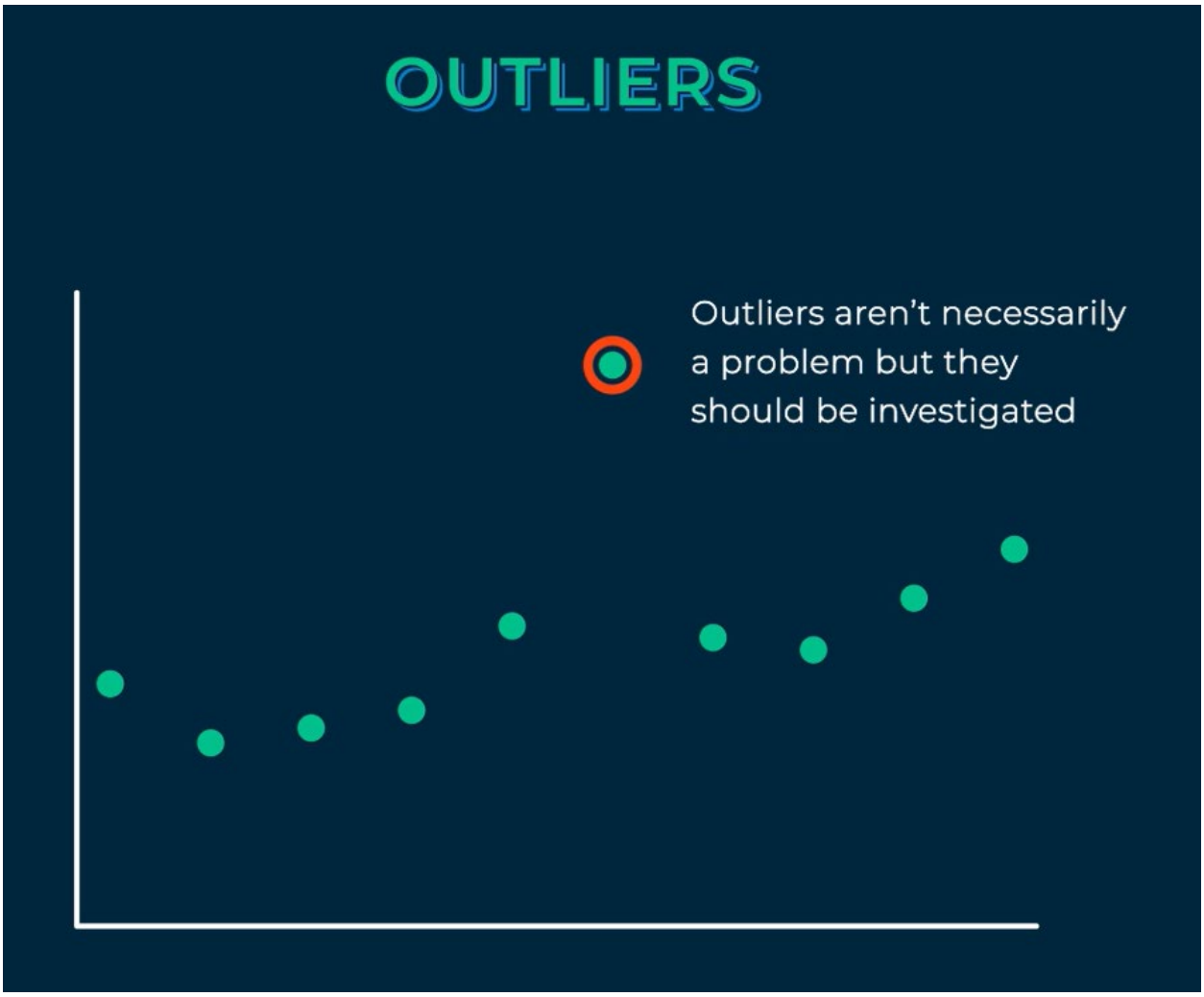
- True
- False

DATA REQUIREMENTS

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

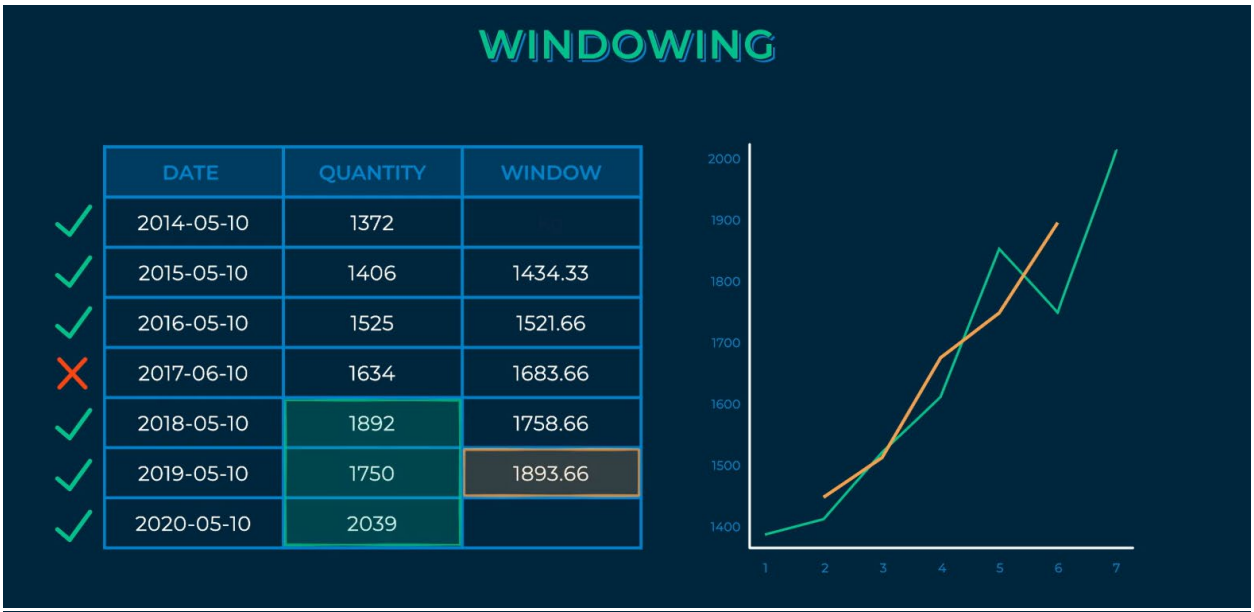
- Must be Numeric
- Single Column
- Consistent Units
- Date = Ascending Order



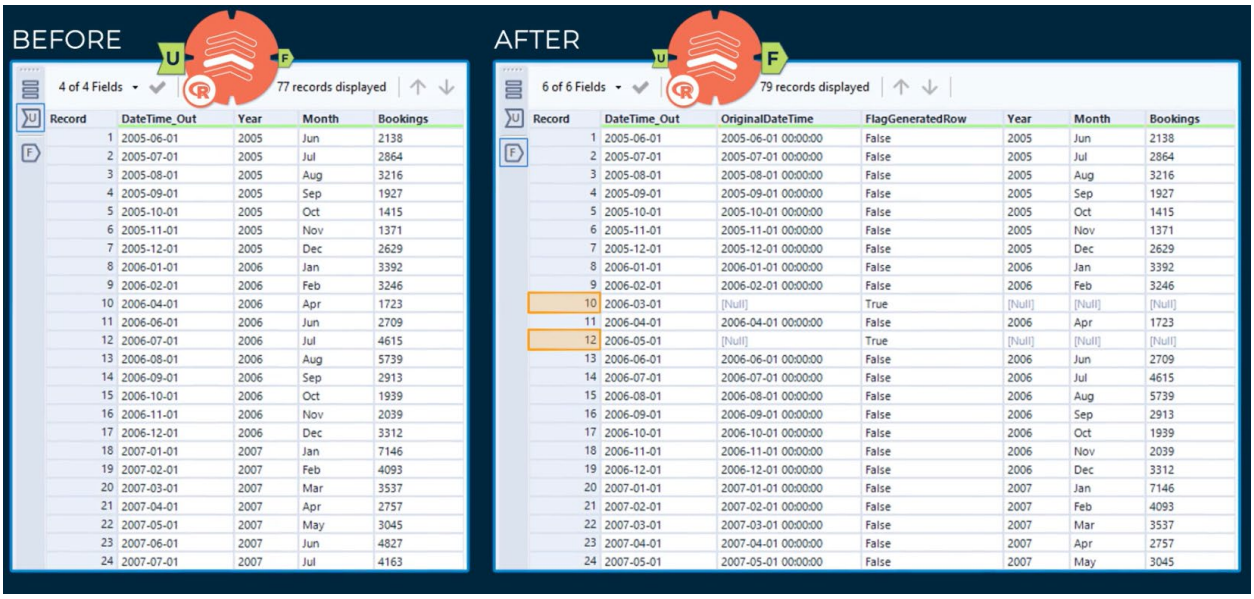
An example of capping outliers:

OUTLIERS

```
If [forecast] > 850 then 850
elseif [forecast] < 150 then 150
else [forecast] Endif
```



May use the TS Filler tool to replace missing values; an example is shown below. ***[Null]*** will be shown in the newly created rows.



BEFORE

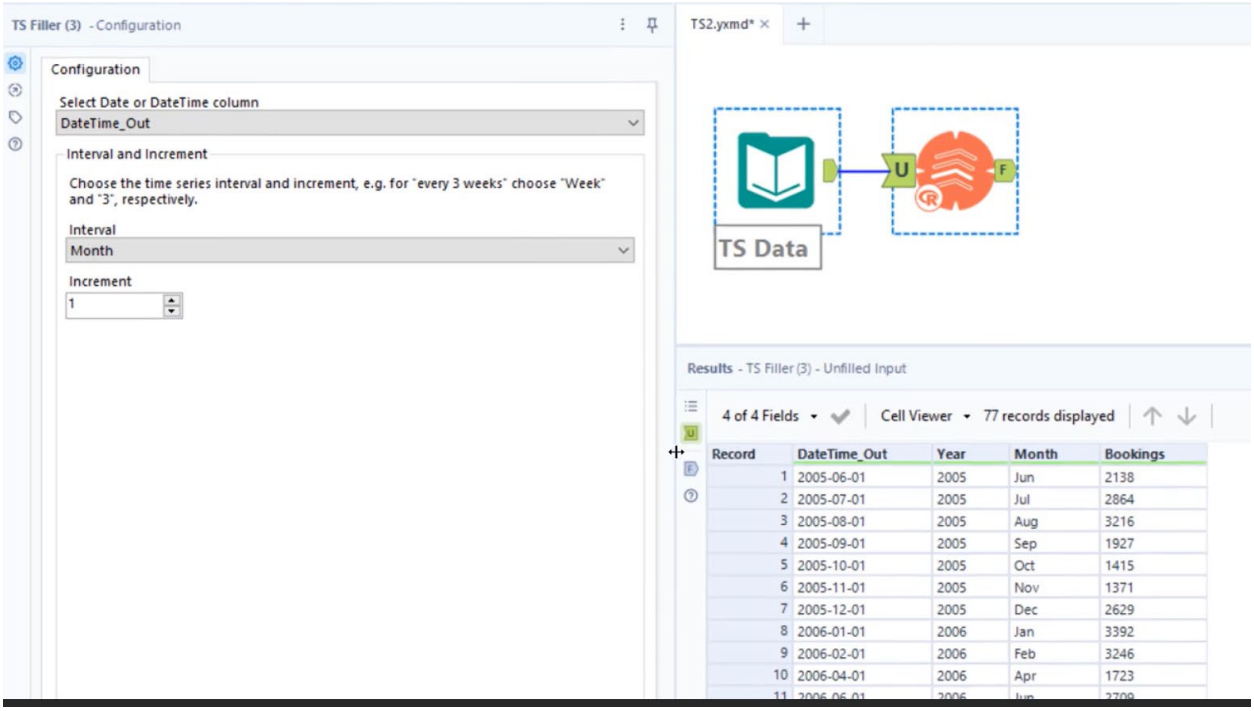
4 of 4 Fields — 77 records displayed

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

AFTER

6 of 6 Fields — 79 records displayed

Record	DateTime_Out	OriginalDateTime	FlagGeneratedRow	Year	Month	Bookings
1	2005-06-01	2005-06-01 00:00:00	False	2005	Jun	2138
2	2005-07-01	2005-07-01 00:00:00	False	2005	Jul	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



TS Filler (3) - Configuration

Configuration

Select Date or DateTime column

DateTime_Out

Interval and Increment

Choose the time series interval and increment, e.g. for "every 3 weeks" choose "Week" and "3", respectively.

Interval

Month

Increment

1

TS2.yxmd*

TS Data

Results - TS Filler (3) - Unfilled Input

4 of 4 Fields — Cell Viewer — 77 records displayed

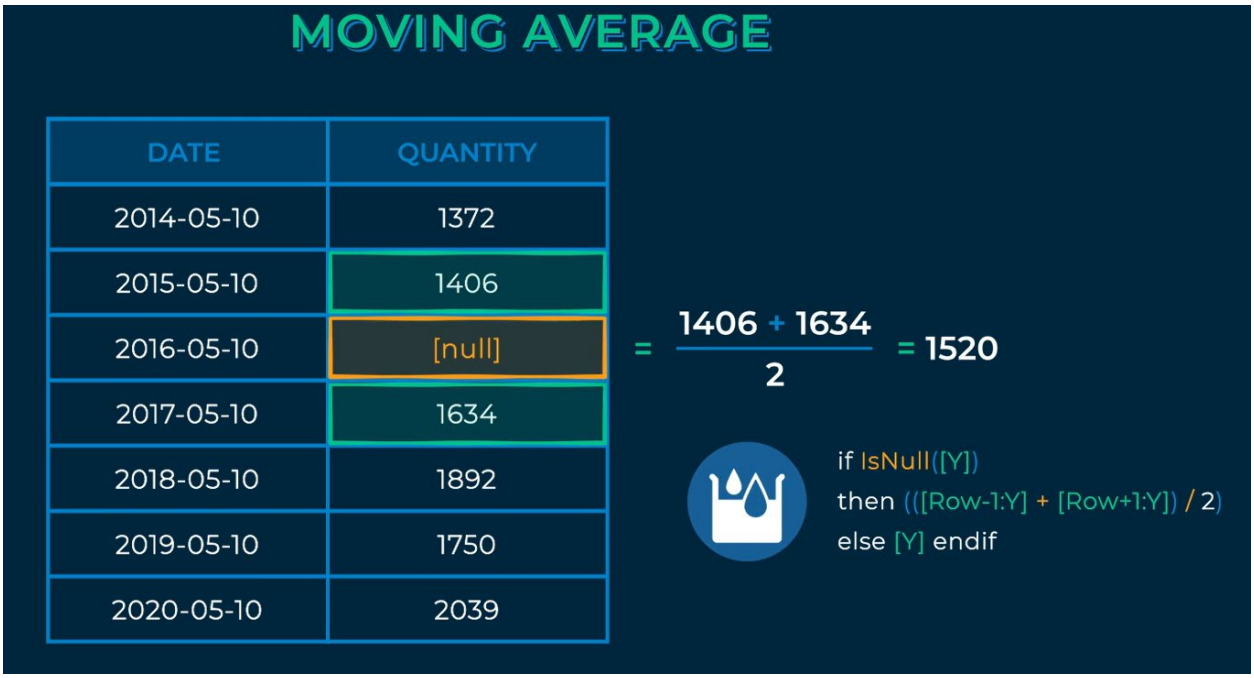
Record	DateTime_Out	Year	Month	Bookings
1	2005-06-01	2005	Jun	2138
2	2005-07-01	2005	Jul	2864
3	2005-08-01	2005	Aug	3216
4	2005-09-01	2005	Sep	1927
5	2005-10-01	2005	Oct	1415
6	2005-11-01	2005	Nov	1371
7	2005-12-01	2005	Dec	2629
8	2006-01-01	2006	Jan	3392
9	2006-02-01	2006	Feb	3246
10	2006-04-01	2006	Apr	1723
11	2006-06-01	2006	Jun	2709

Results - TS Filler (3) - Filled Output

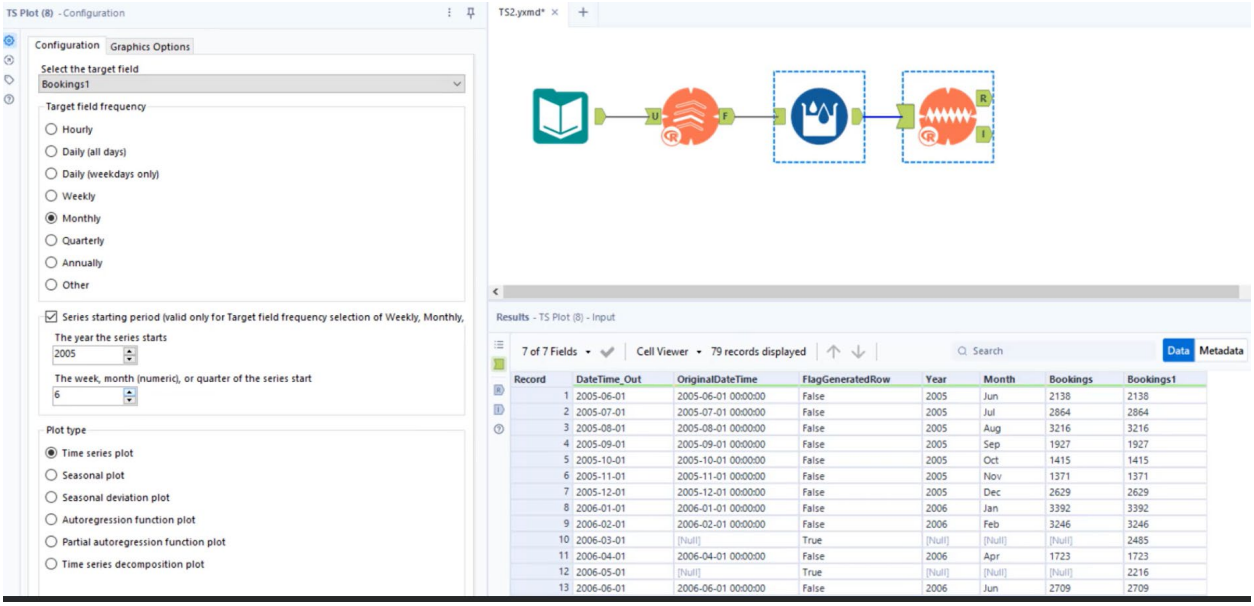
6 of 6 Fields | Cell Viewer | 79 records displayed | Search

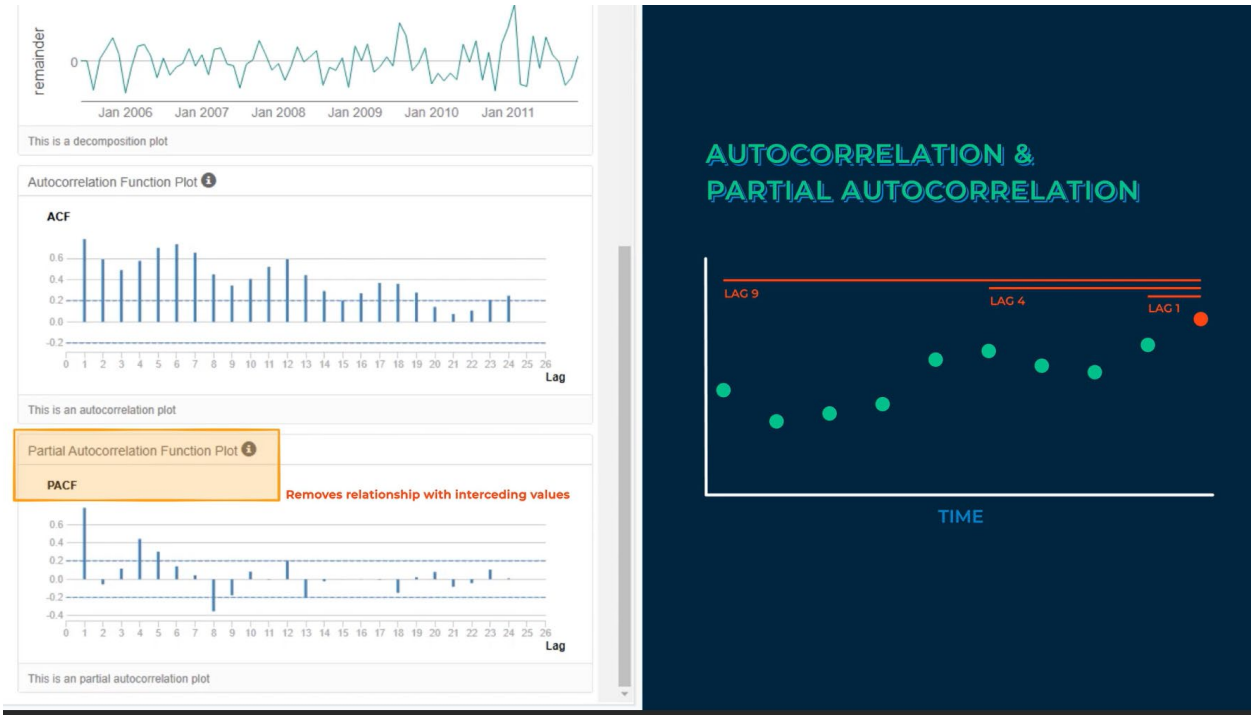
Record	DateTime_Out	OriginalDateTime	FlagGeneratedRow	Year	Month	Bookings
1	2005-06-01	2005-06-01 00:00:00	False	2005	Jun	2138
2	2005-07-01	2005-07-01 00:00:00	False	2005	Jul	2864
3	2005-08-01	2005-08-01 00:00:00	False	2005	Aug	3216
4	2005-09-01	2005-09-01 00:00:00	False	2005	Sep	1927
5	2005-10-01	2005-10-01 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

One method of filling in the missing values is to calculate the moving average, as shown below.



You can then investigate your data by using a TS Plot tool, as shown below.


TS Plot (8) - Configuration
Configuration
Graphics Options
Select the target field
Bookings1
Target field frequency
Hourly
Daily (all days)
Daily (weekdays only)
Weekly
Monthly
Quarterly
Annually
Other
Series starting period (valid only for Target field frequency selection of Weekly, Monthly,
The year the series starts
2005
The week, month (numeric), or quarter of the series start
6
Plot type
Time series plot
Seasonal plot
Seasonal deviation plot
Autoregression function plot
Partial autoregression function plot
Time series decomposition plot
TS2.yxmd*
Results - TS Plot (8) - Input
7 of 7 Fields
Cell Viewer
79 records displayed
Search
Data
Metadata
Record
DateTime_Out
OriginalDateTime
FlagGeneratedRow
Year
Month
Bookings
Bookings1
1 2005-06-01 2005-06-01 00:00:00 False 2005 Jun 2138 2138
2 2005-07-01 2005-07-01 00:00:00 False 2005 Jul 2864 2864
3 2005-08-01 2005-08-01 00:00:00 False 2005 Aug 3216 3216
4 2005-09-01 2005-09-01 00:00:00 False 2005 Sep 1927 1927
5 2005-10-01 2005-10-01 00:00:00 False 2005 Oct 1415 1415
6 2005-11-01 2005-11-01 00:00:00 False 2005 Nov 1371 1371
7 2005-12-01 2005-12-01 00:00:00 False 2005 Dec 2629 2629
8 2006-01-01 2006-01-01 00:00:00 False 2006 Jan 3392 3392
9 2006-02-01 2006-02-01 00:00:00 False 2006 Feb 3246 3246
10 2006-03-01 [Null] True [Null] [Null] [Null] 2485
11 2006-04-01 2006-04-01 00:00:00 False 2006 Apr 1723 1723
12 2006-05-01 [Null] True [Null] [Null] [Null] 2216
13 2006-06-01 2006-06-01 00:00:00 False 2006 Jun 2709 2709


remainder
0
Jan 2006
Jan 2007
Jan 2008
Jan 2009
Jan 2010
Jan 2011
This is a decomposition plot
Autocorrelation Function Plot
ACF
Lag
This is an autocorrelation plot
Partial Autocorrelation Function Plot
PACF
Removes relationship with interceding values
Lag
This is an partial autocorrelation plot
AUTOCORRELATION &
PARTIAL AUTOCORRELATION
LAG 9
LAG 4
LAG 1
TIME

Question 1

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

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

Question 2

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

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

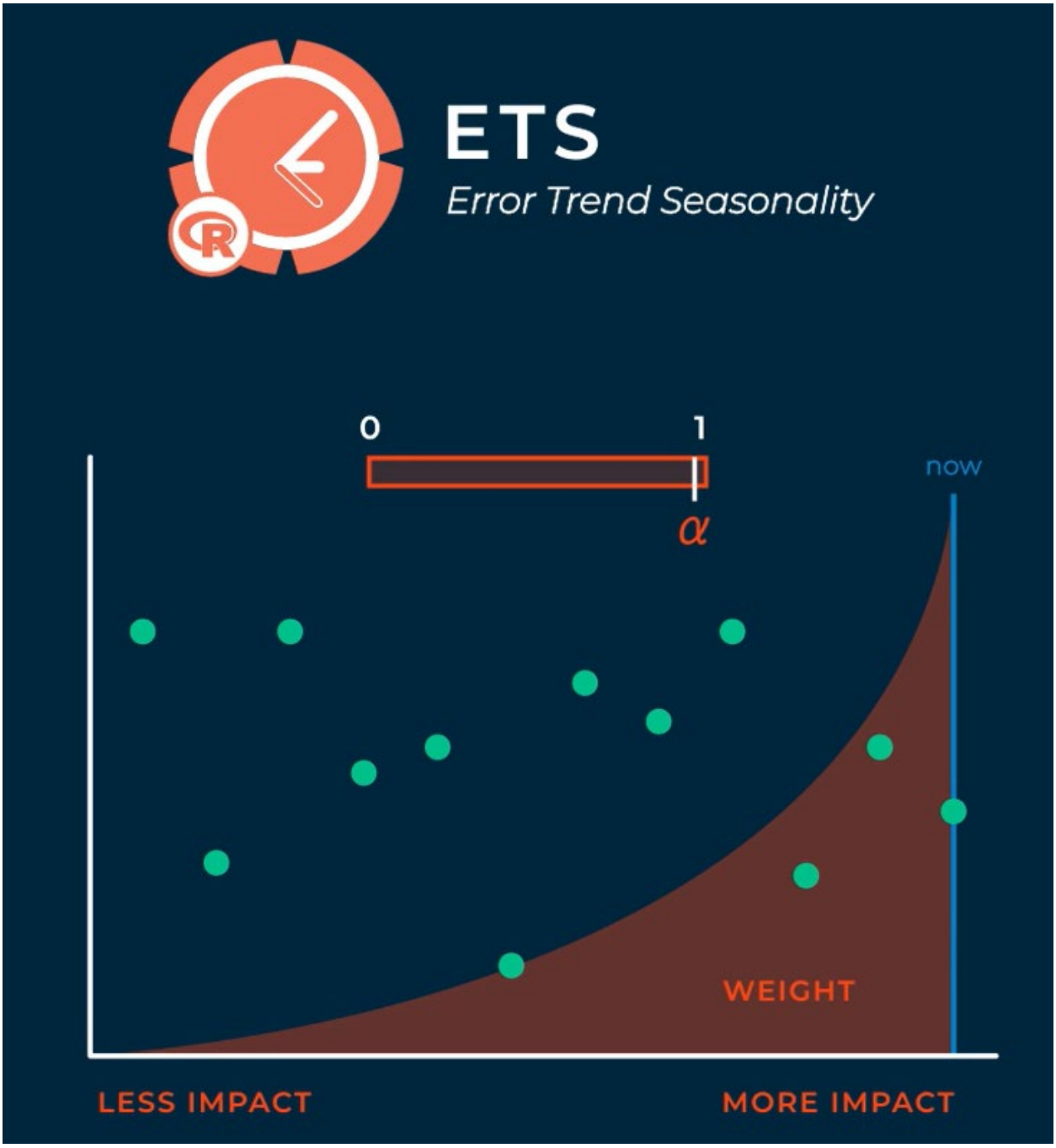
Question 3

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

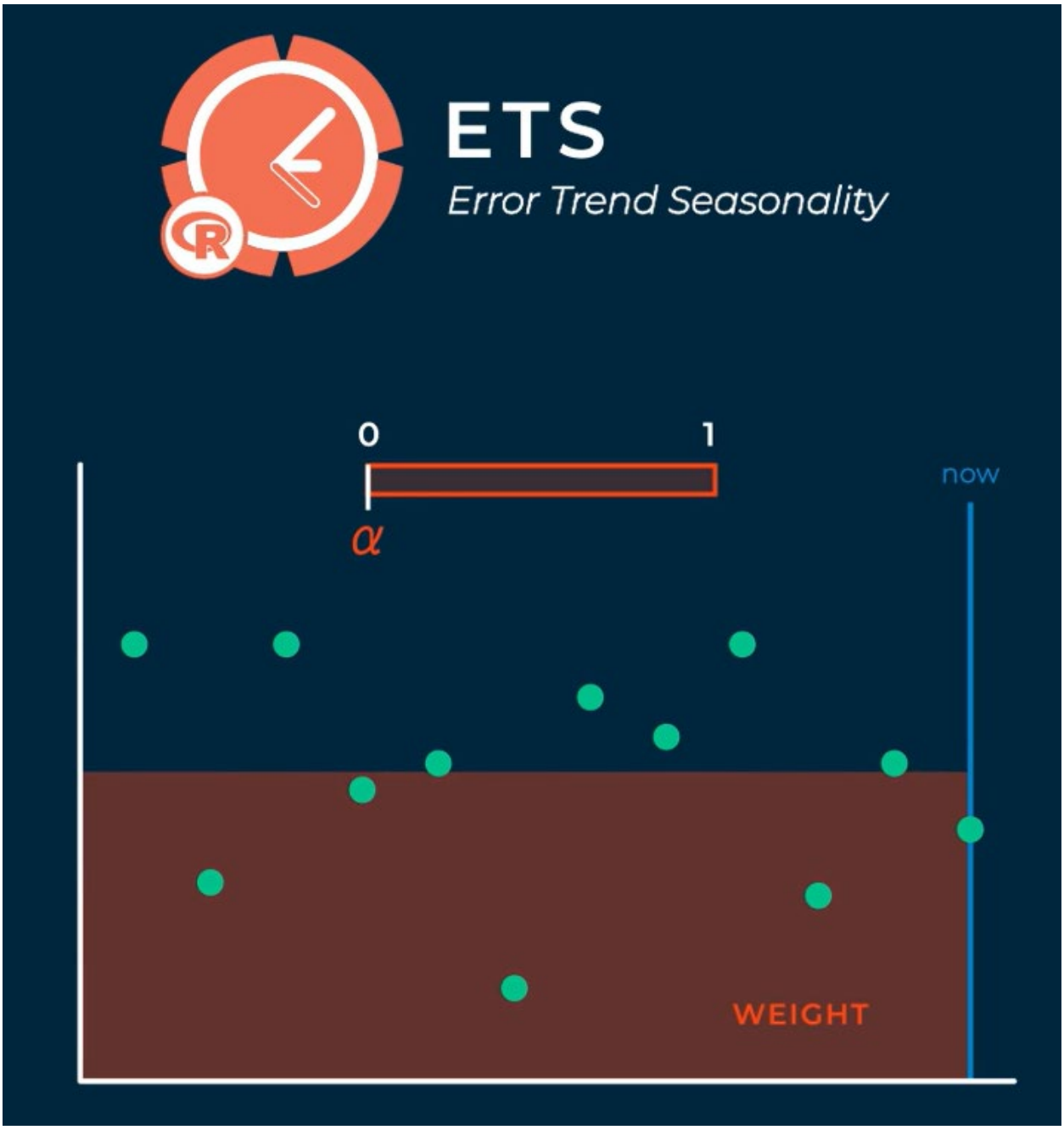
- [ ] A prediction for the next day
- [x] A prediction for the next week
- [x] A prediction for the next month

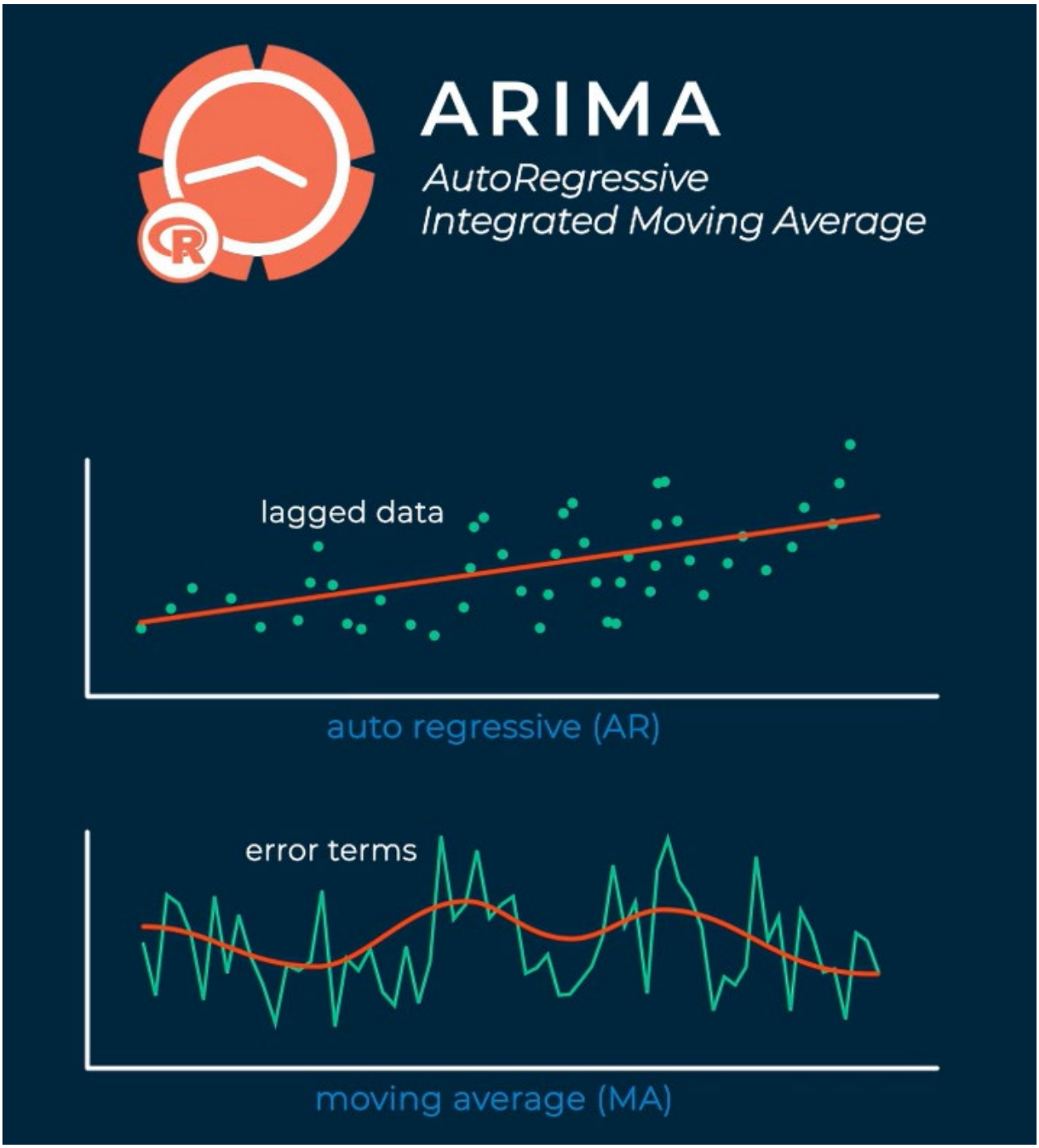
	Year	Week_Of	Total Consumption EUR	Total Price EUR	Total Consumption SIB	Total Price SIB
1	2006	2006-09-01 00:00:00	11460707	58252.76	3146358	58252.76
2	2006	2006-09-08 00:00:00	11882622	78173.55	3171588	78173.55
3	2006	2006-09-15 00:00:00	12333684	80129.9	3178116	80129.9
4	2006	2006-09-22 00:00:00	12485116	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

64. ETS & ARIMA



With ETS, if alpha is set to 0, then all data points are weighted equally.


ETS
Error Trend Seasonality
0
1
α
now
WEIGHT


ARIMA
AutoRegressive
Integrated Moving Average
lagged data
auto regressive (AR)
error terms
moving average (MA)

INTEGRATED

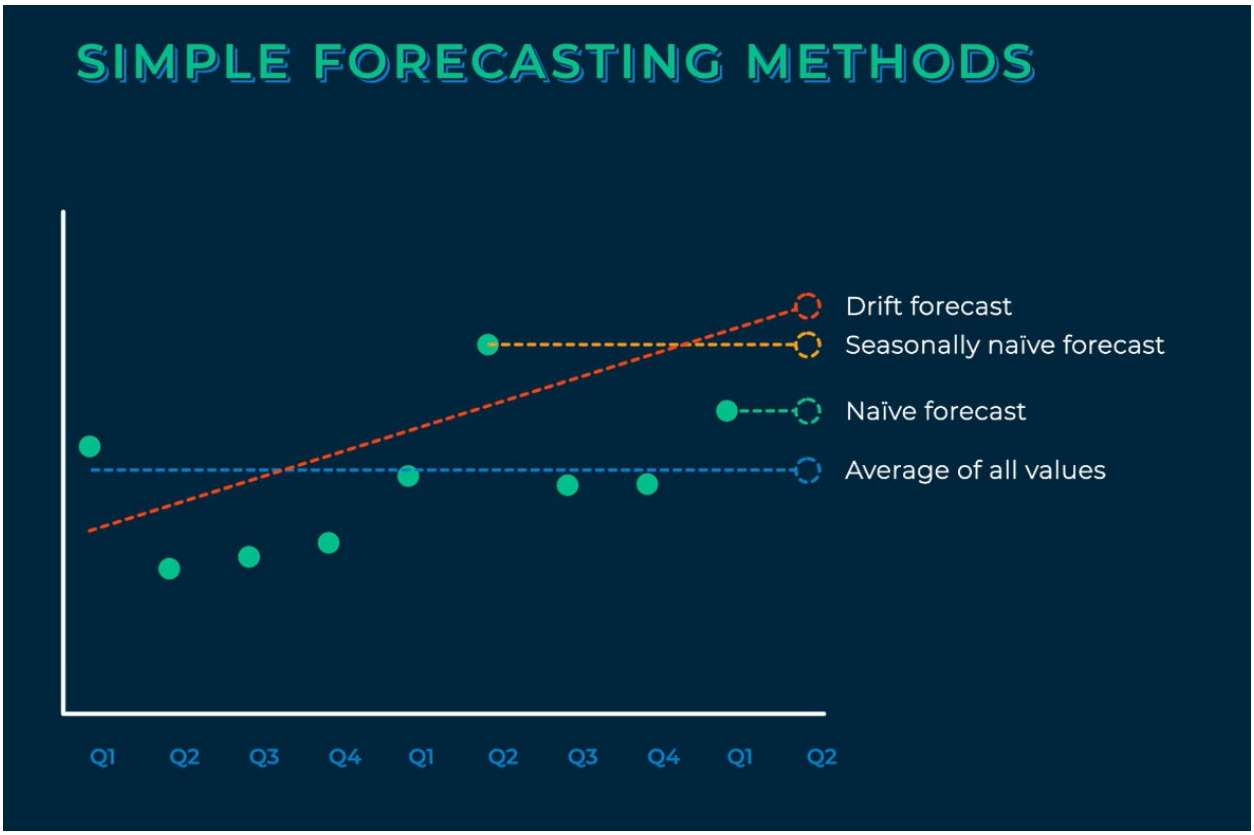
Level of Differencing

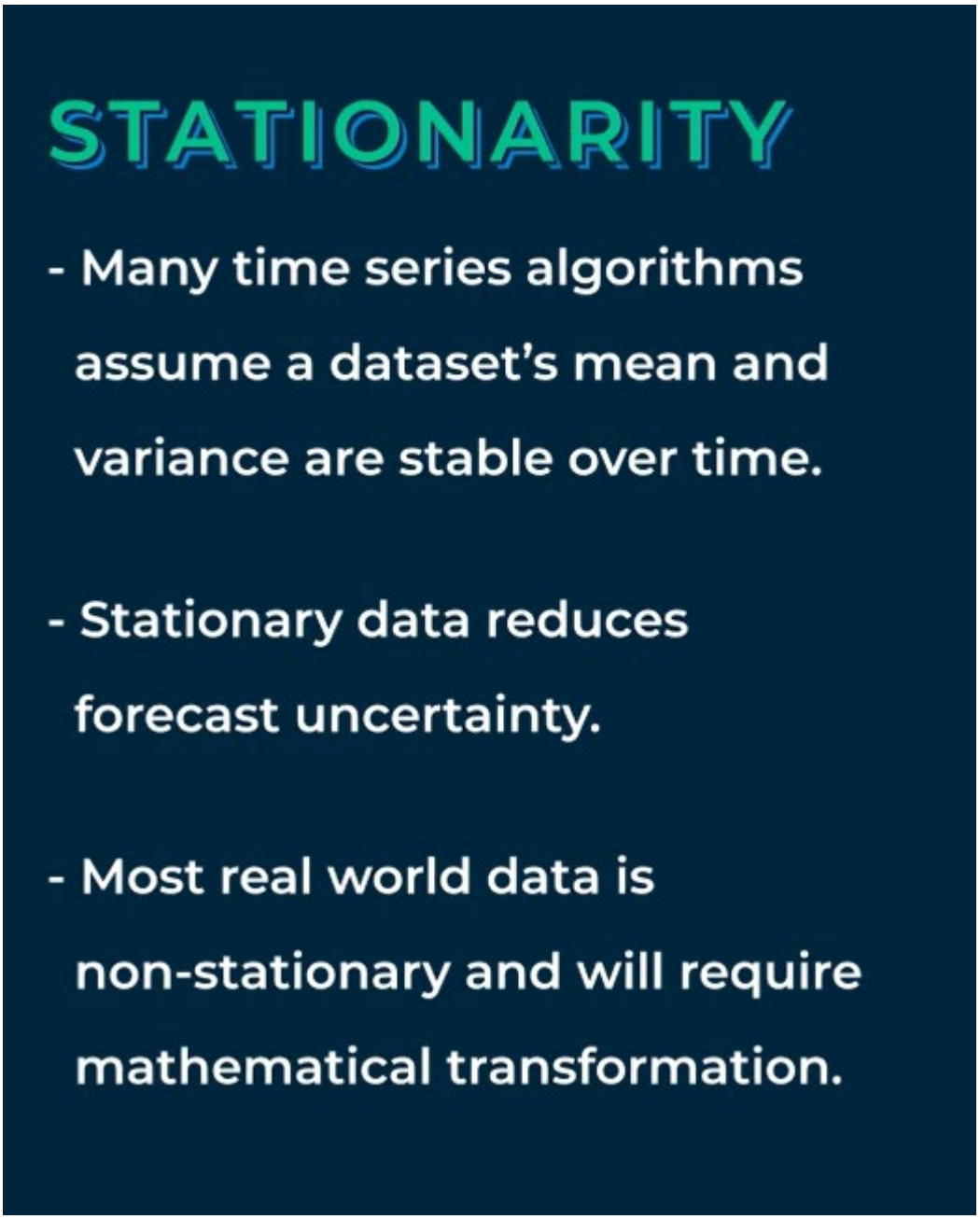
DIFFERENCING

Makes Data Stationary

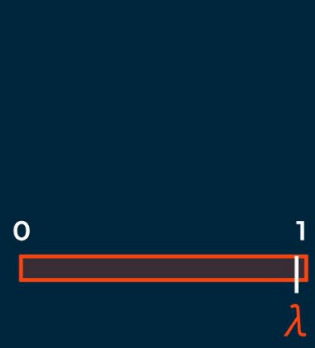
AUTOCORRELATION

Similarity of Observations at Various Lags


SIMPLE FORECASTING METHODS
Drift forecast
Seasonally naïve forecast
Naïve forecast
Average of all values
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
Q1
Q2



A Box-Cox transformation is used to stabilize the variance of a dataset. This allows you to change the value of lambda. With lambda equal to 1, the data are shifted, and the shape is unchanged. With lambda equal to 0, the result is equivalent to taking the natural log of the data. All other values are an exponential power function.

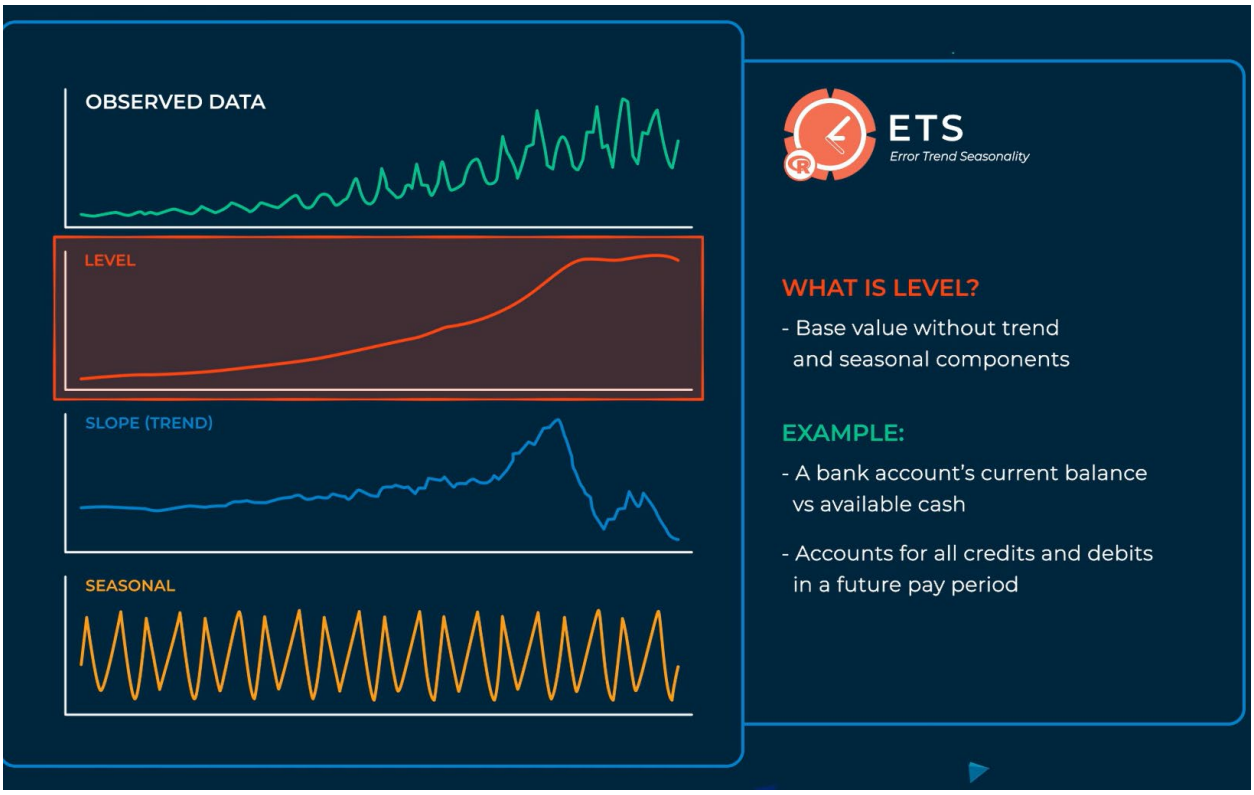


BOX-COX
TRANSFORMATION

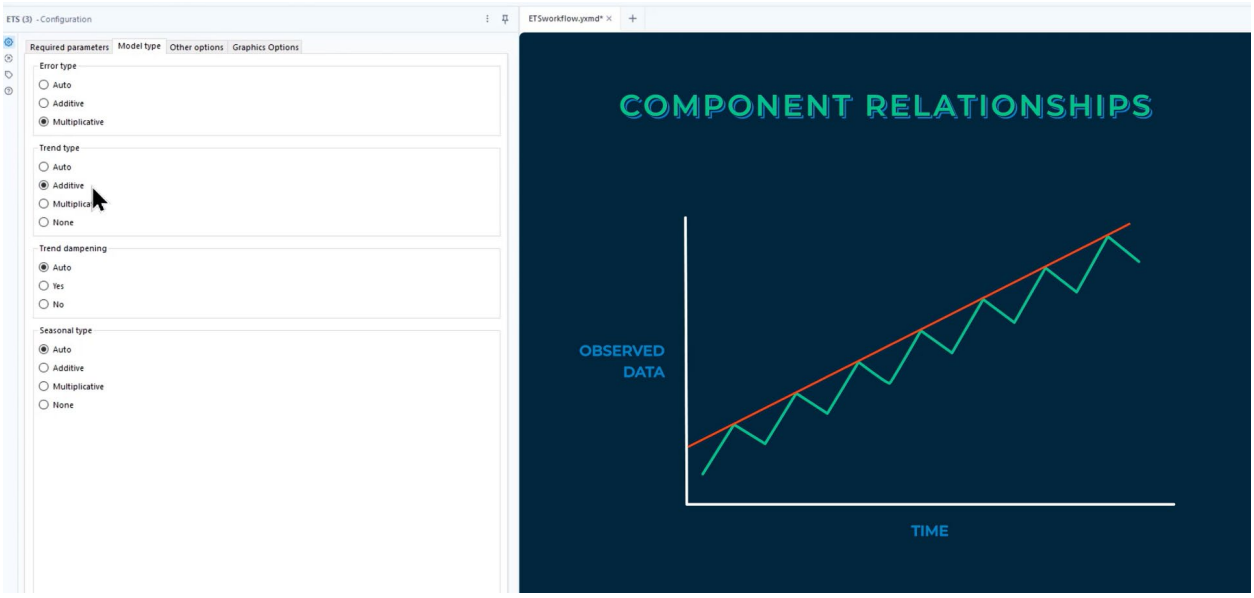
BOX-COX TRANSFORMATION

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

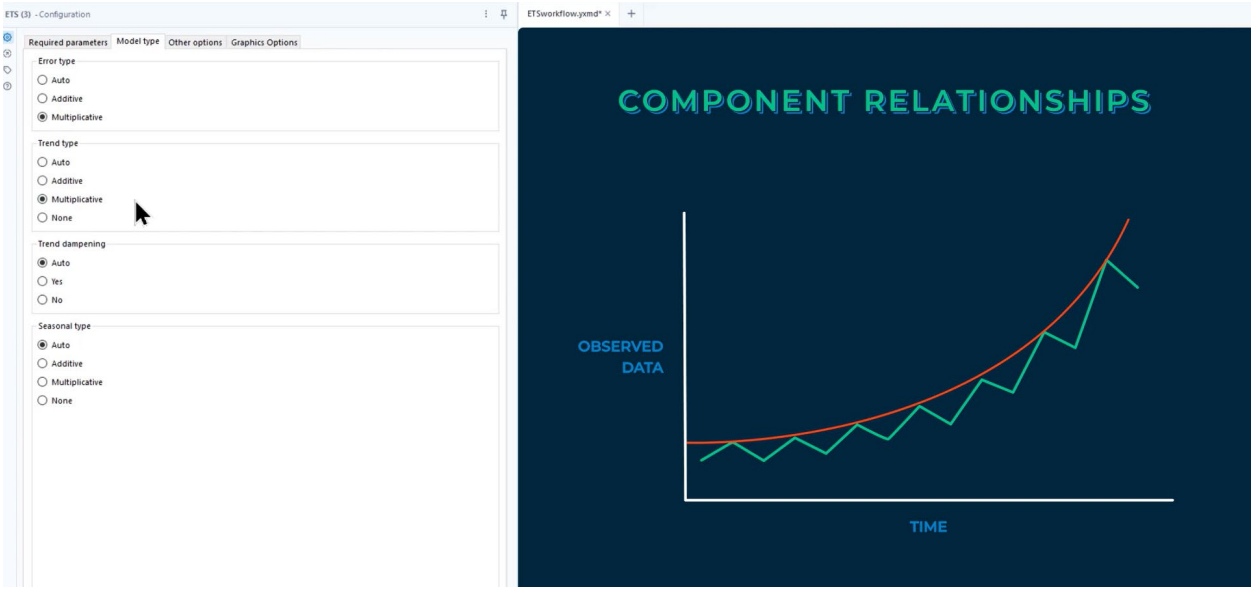
TS algorithms will adjust mean
stabilization automatically.

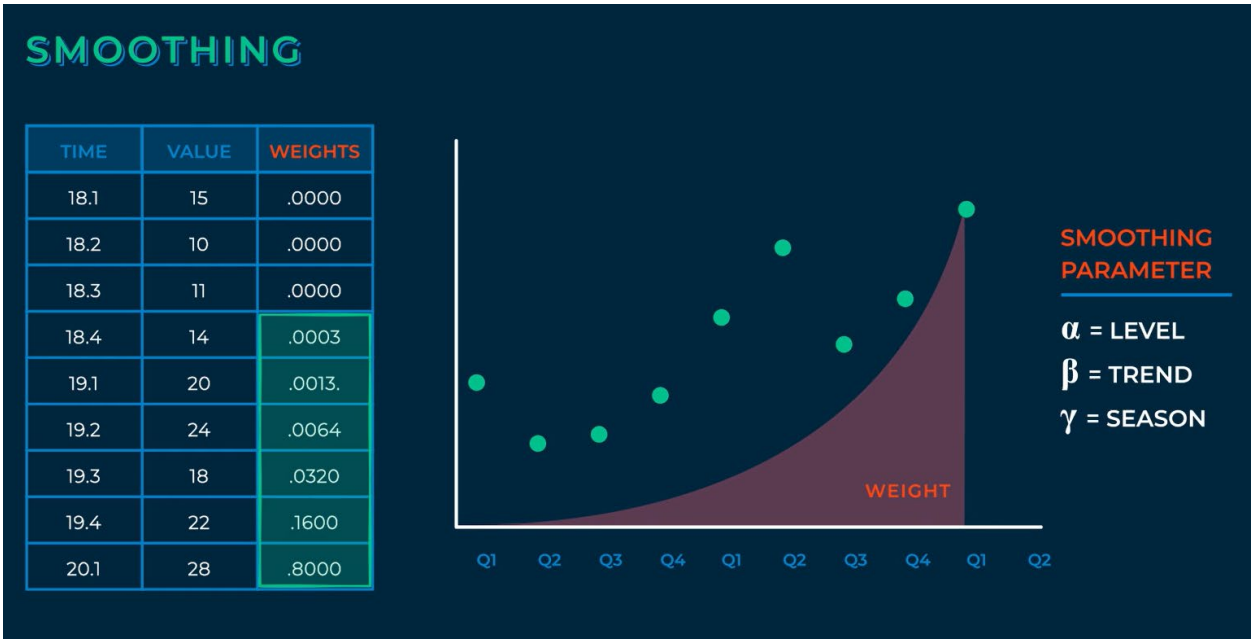


If the general shape of the data is linear, an additive trend can be assumed.

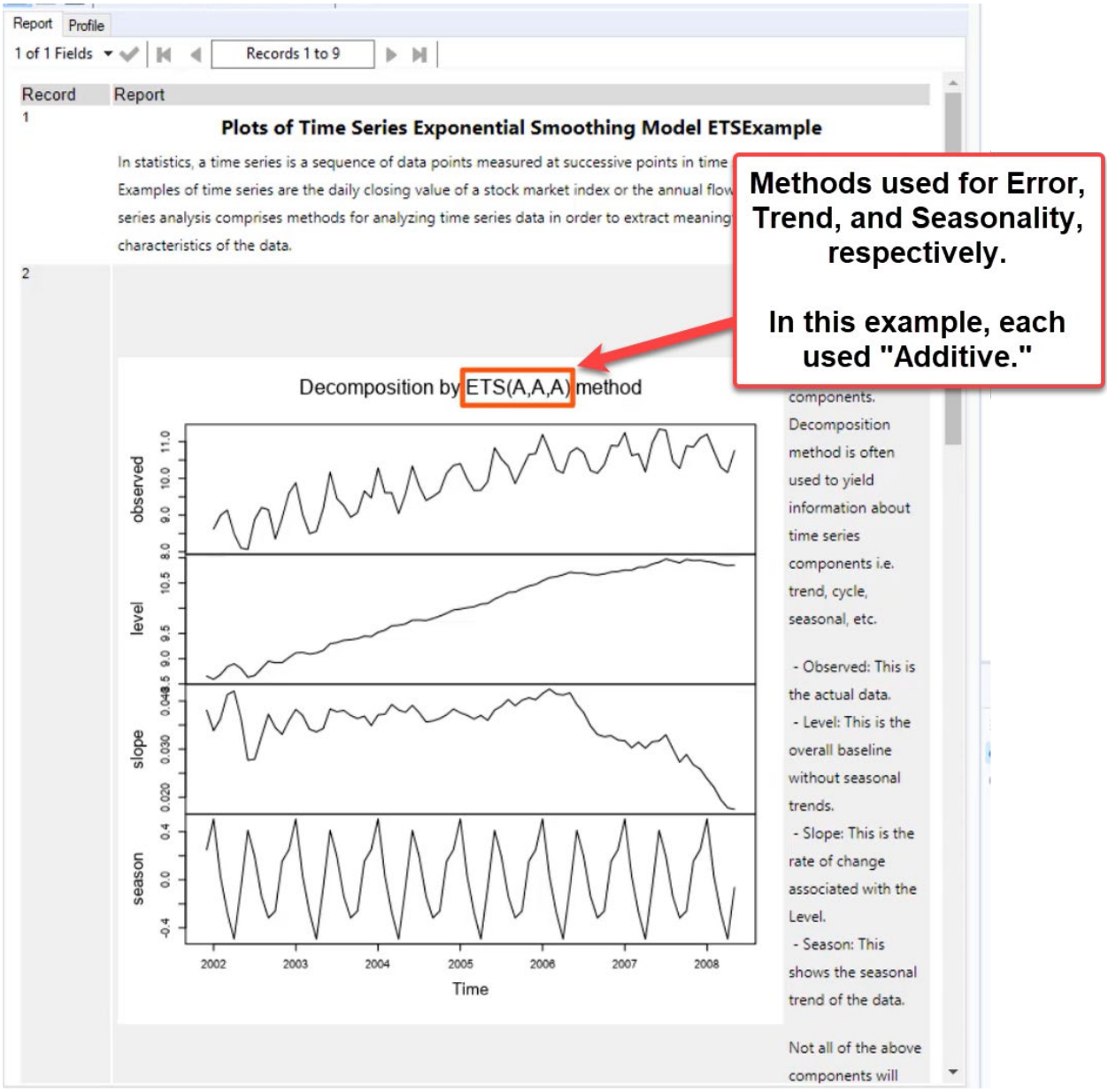


An exponential shape can indicate multiplicative trending.





Phi is dampening.



Trend may include a subscript to indicate dampening.

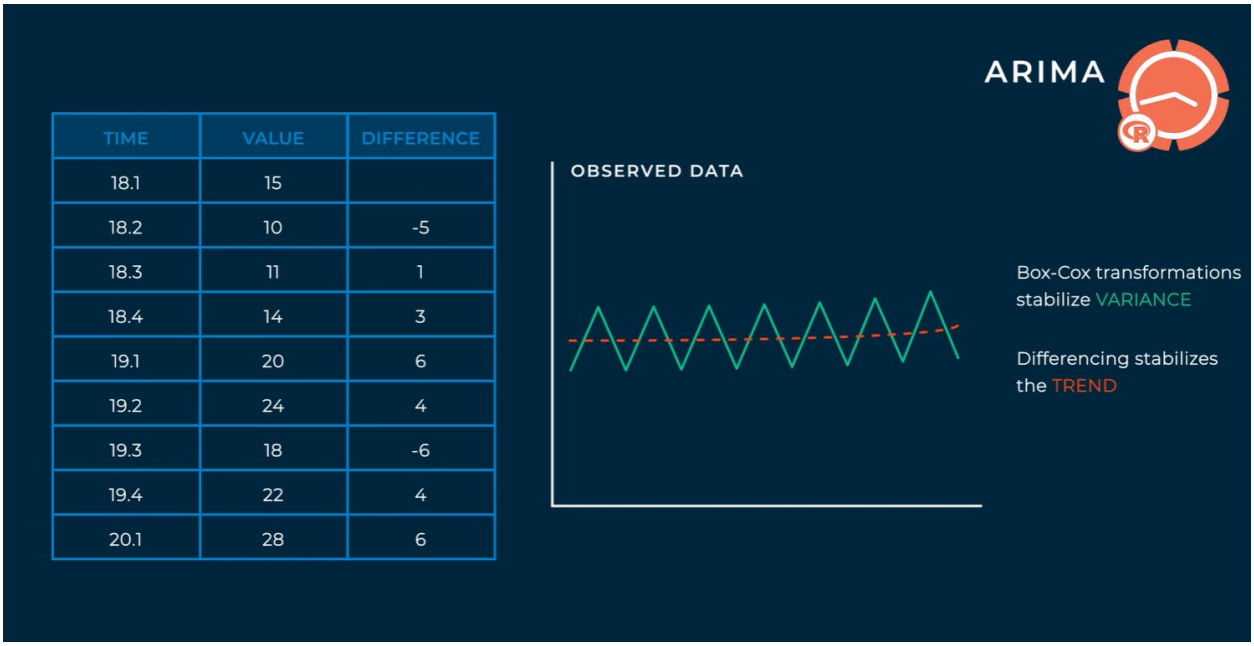
A = additive

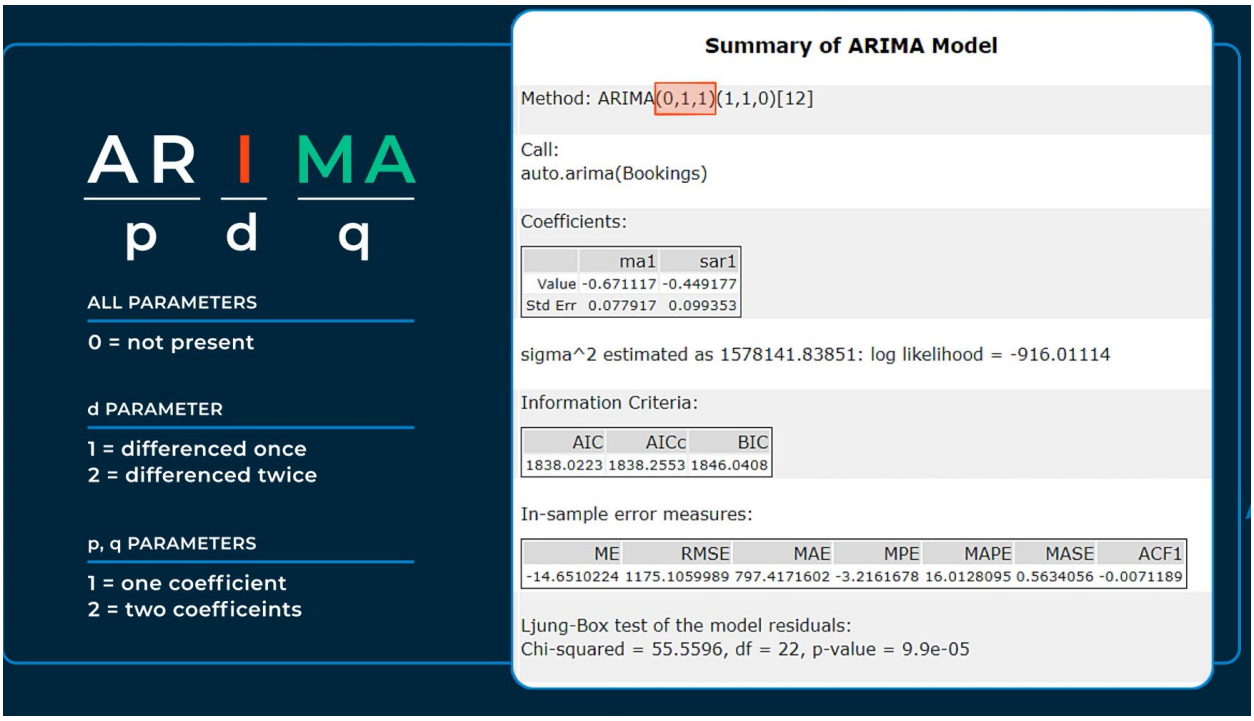
M = multiplicative

N = none

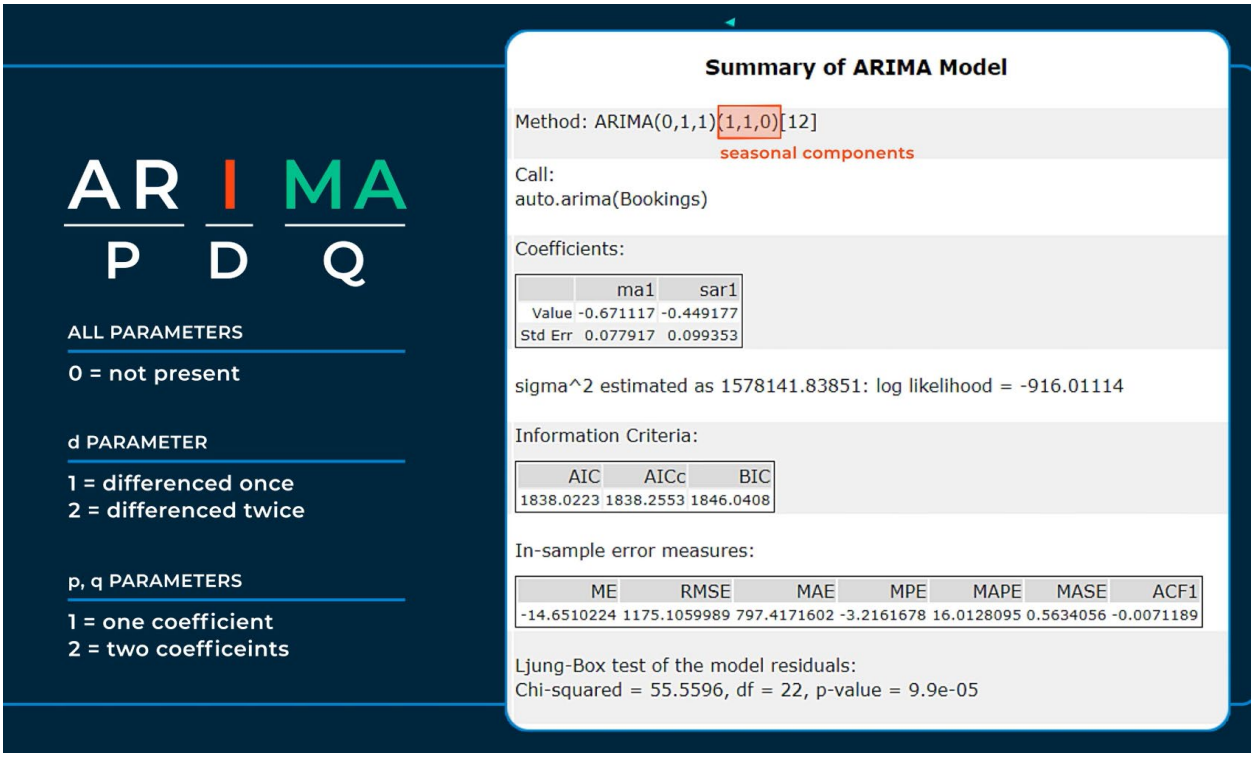
d = dampening

ARIMA

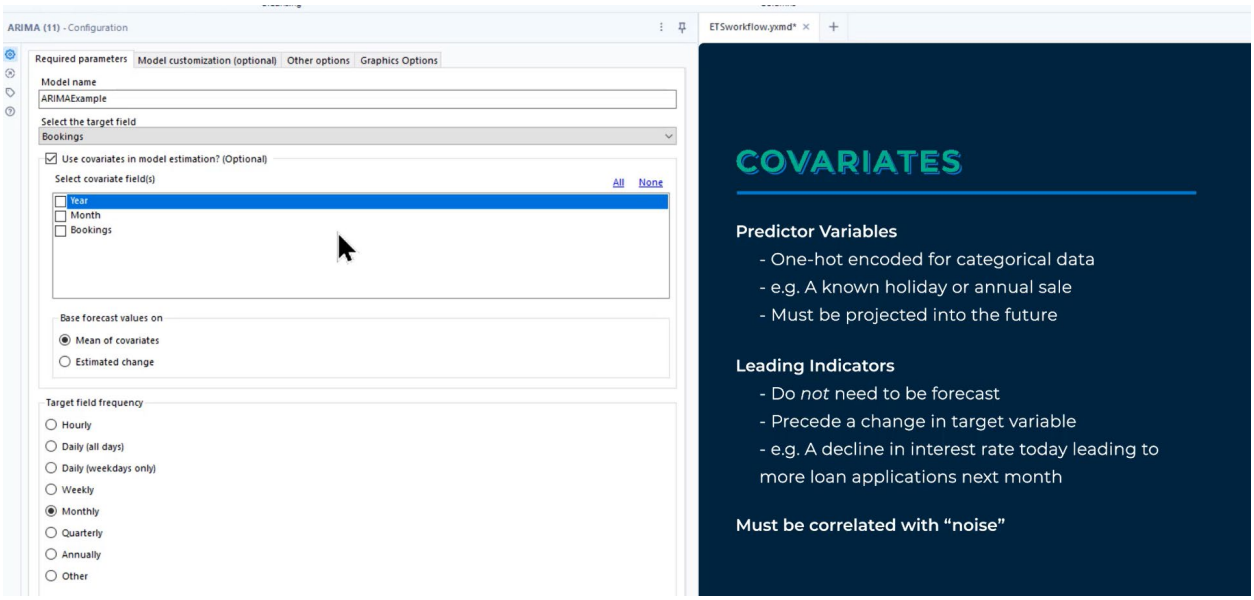




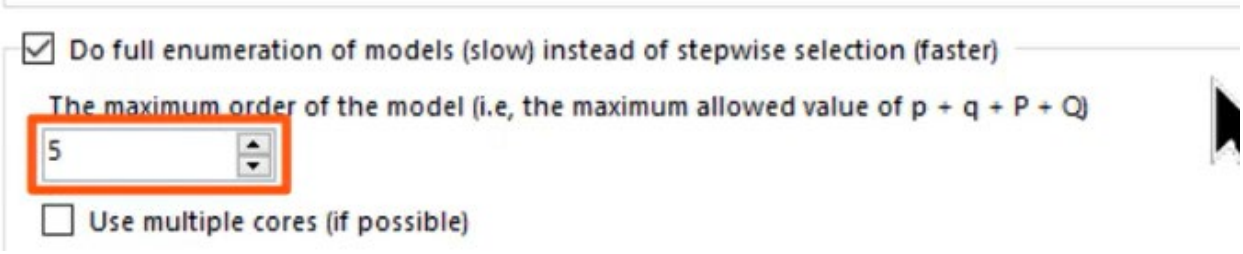
Seasonal components used capitalized P, D, and Q.



Periods in the seasonal cycle:



The "Do full enumeration of models (slow) instead of stepwise selection (faster)" check box will increase the time required to run the workflow but will allow the algorithm to compare more models when selecting the best performer.



The "Allow drift" check box allows the final regression to have a constant value (similar to a y-intercept).

Results - Browse (14) - Input

LJUNG-BOX TEXT

Determines if residuals aurocorrelations are non-zero

- P-value **> .05 =** residuals not autocorrelated (good)
- A very low p-value means the model may have missed some relationships

Question 1

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

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

Question 2

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

- No seasonal AR term
- No MA term
- None of these
- No seasonal MA term
- No AR term

Question 3

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

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

Question 4

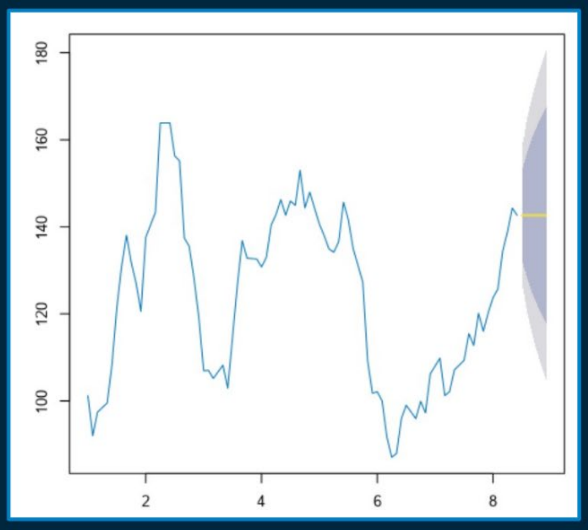
What is the purpose of a Box-Cox transformation?

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

Question 5

Which type of simple forecast is this?

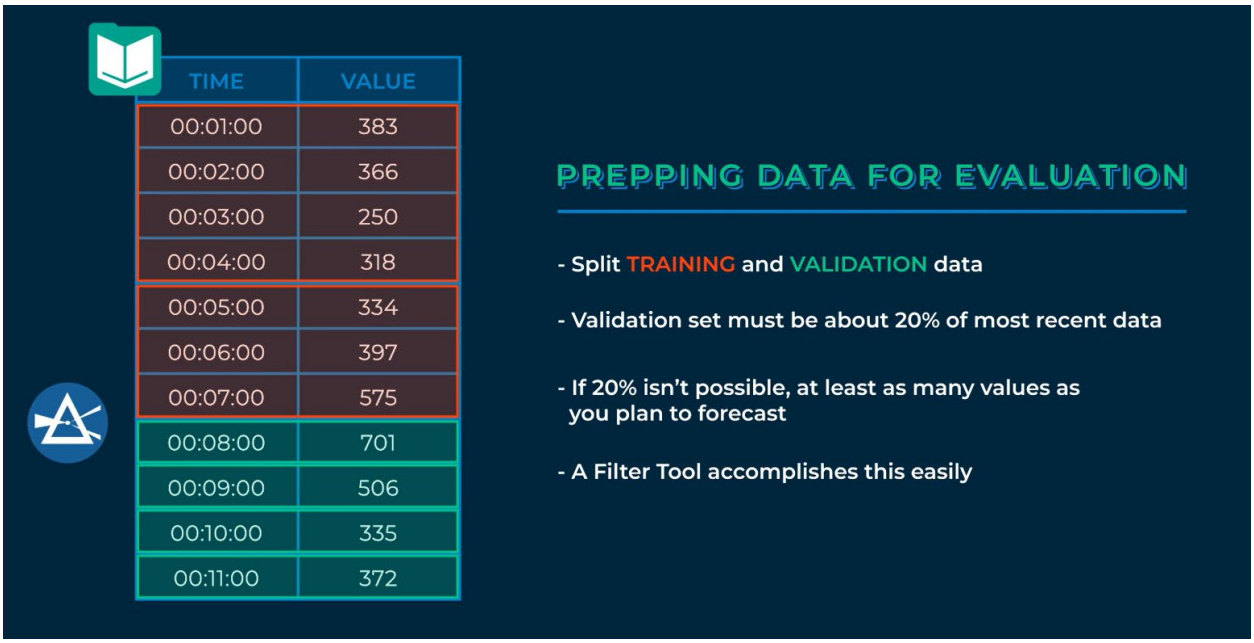
- Dampened Drift
- Drift
- Naïve
- Seasonal Naïve

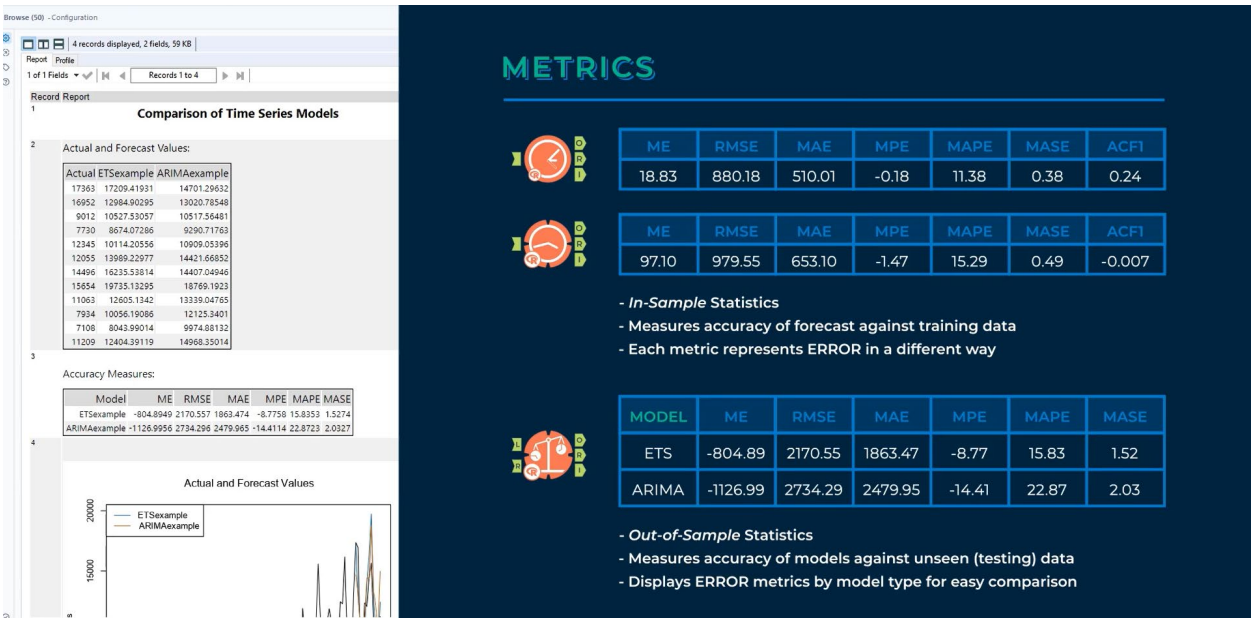


Question 6

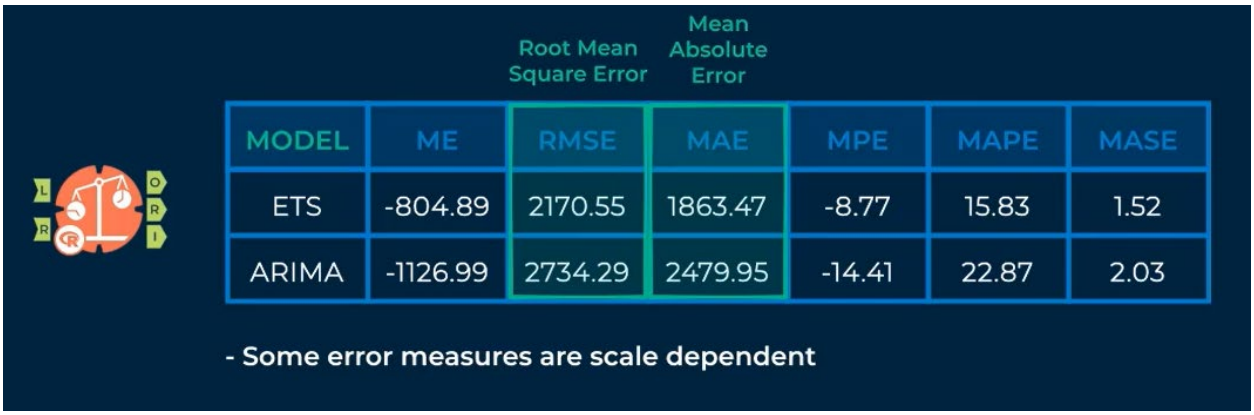
What does an alpha close to 1 indicate in an ETS model?

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

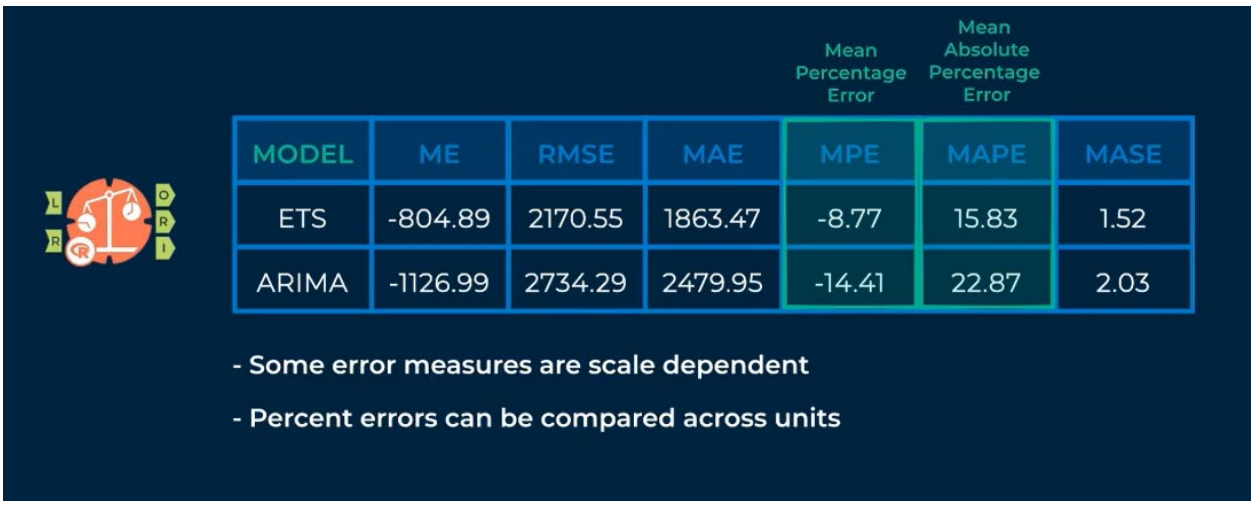




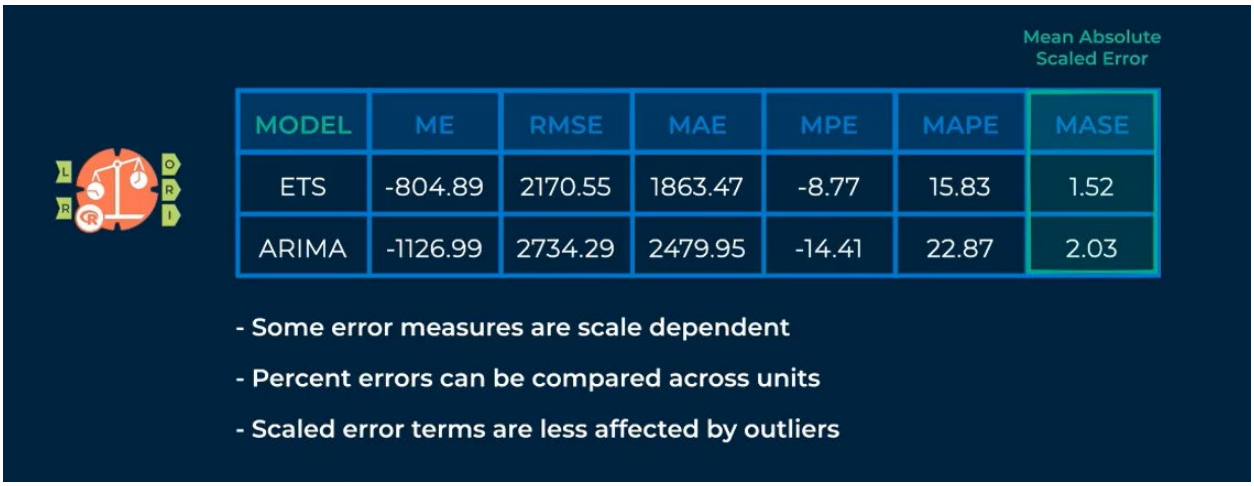
Scale dependent measurements such as RMSE and MAE should not be compared to models with different units.



MPE and MAPE can be compared between models.



MASE is an alternative to percentage errors when comparing models created with different units.

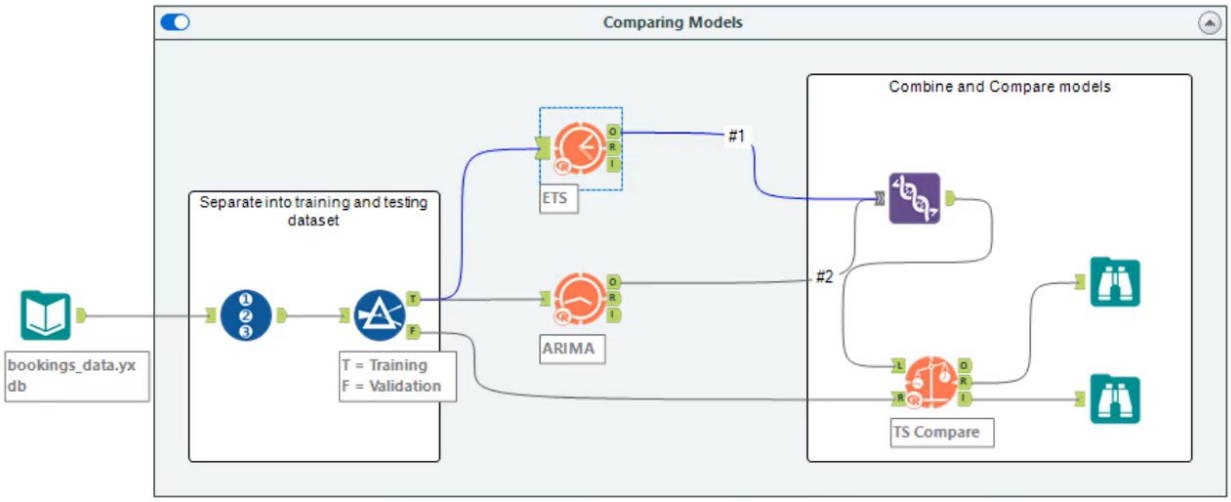


Values with the lowest absolute value indicate better performance.

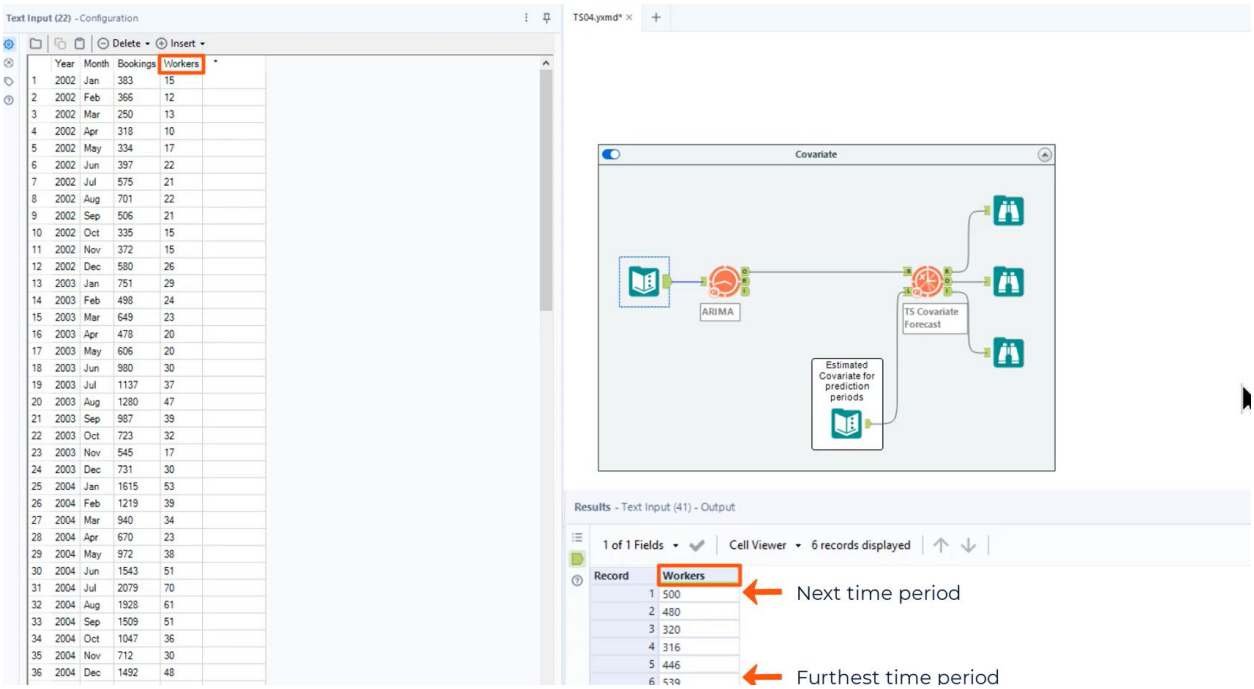
MODEL	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-804.89	2170.55	1863.47	-8.77	15.83	1.52
ARIMA	-1126.99	2734.29	2479.95	-14.41	22.87	2.03

- Some error measures are scale dependent
- Percent errors can be compared across units
- Scaled error terms are less affected by outliers
- Values approaching zero indicate better performance

After selecting a model, a forecast is necessary. To make a forecast, use all of the data available (which includes the most recent values). Models compared may use only a training dataset (and compared using a validation dataset). See the screenshot below for an example.



If using an ARIMA model with a covariate, the TS Covariate Forecast tool must be used. The data that go into the L input anchor should have the next time period at the top and the furthest time period at the bottom and the column header needs to match a column header in the original dataset. See the screenshot below for an example.



TIME SERIES CONSIDERATIONS

- Data investigation & cleanup are particularly important
- Results may be nonsensical without data prep
- Follow best practices for each model type
- Time series model lifespans are short
- Upstream dataprep can be applied to updated datasets to create new models quickly

Question 1

How is MAPE different from RMSE?

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

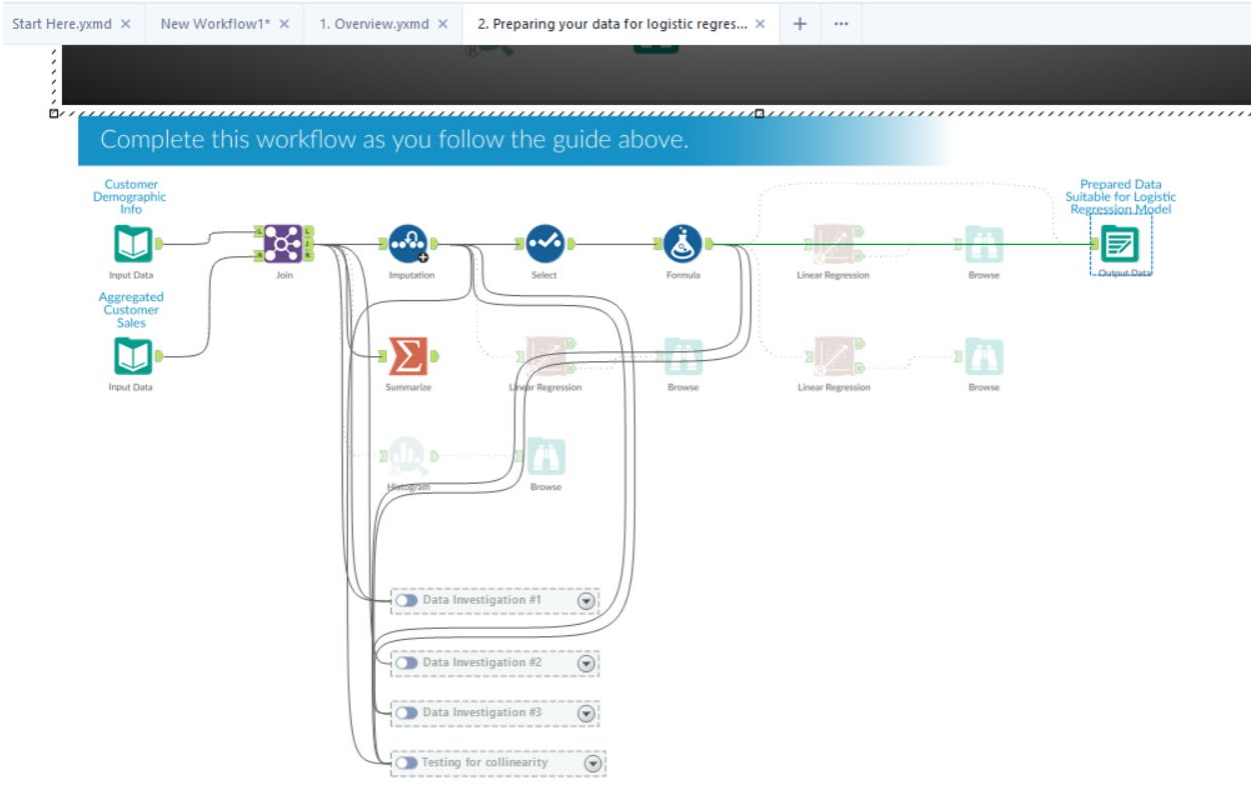
- One is a prediction and the other is a forecast.
- One is an in-sample statistic and the other measures performance on unseen data.
- One is calculated using an estimated standard deviation and the other uses the computed standard deviation for the dataset.
- There is no difference.

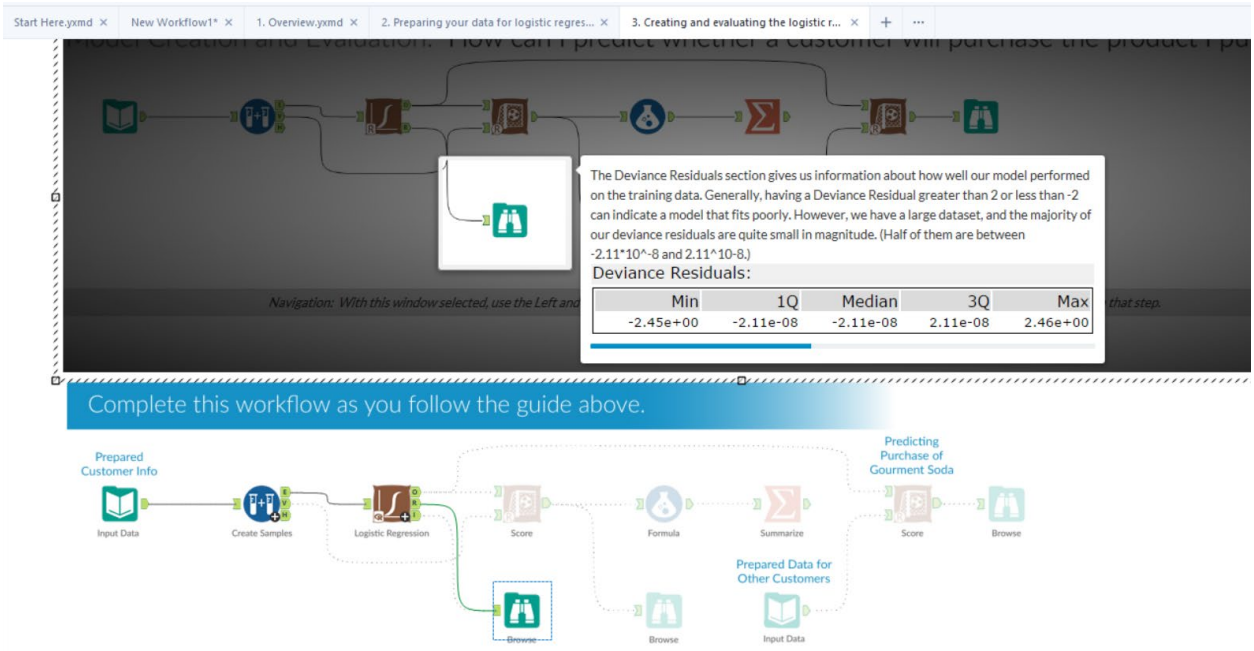
Question 3

Why can't you forecast from the best algorithm connected to the TS Compare tool?

- The TS Compare tool cannot pass along the model object.
- That model object was created using an incomplete dataset.
- You can and should use it.
- Running the model object to both streams will take longer.

65. The Akaike Information Criterion
66. Examples

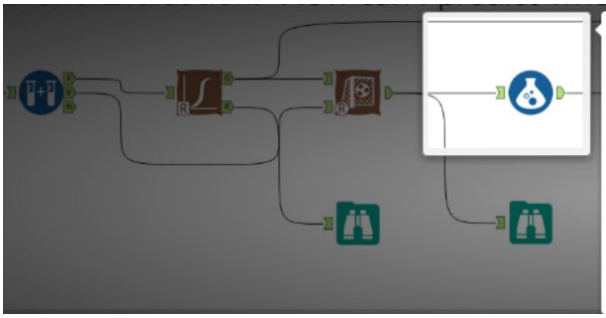




The null model is the model with just an intercept term. That is, the null model predicts the same value for every record, because it does not use any predictor variables. Comparing the null deviance to the residual deviance provides us with information about how well our model outperforms the null model. The McFadden R-squared value summarizes the relative values of these two deviances.

In general, the higher the McFadden R-squared value, the more the current model outperforms the null model. A McFadden R-squared value between .2 and .4 indicates a relatively good model. Thus, our McFadden R-squared value of .9607 is excellent.

```
Null deviance: 5541.2 on 3999 degrees of freedom
Residual deviance: 217.64 on 3995 degrees of freedom
McFadden R-Squared: 0.9607, AIC: 227.6
```



The logic behind the expression is that if the item was purchased and the value from `Score_1` (i.e. Score True) is greater than .5, the prediction is correct. (If the Logistic Regression tool predicts that a record's probability of being true is at least .5, then the tool generates a true prediction for that record.) If the item was purchased and `Score_1` is less than or equal to .5, the model made an incorrect prediction. The logic is switched for cases where the item was not purchased.

Expression:

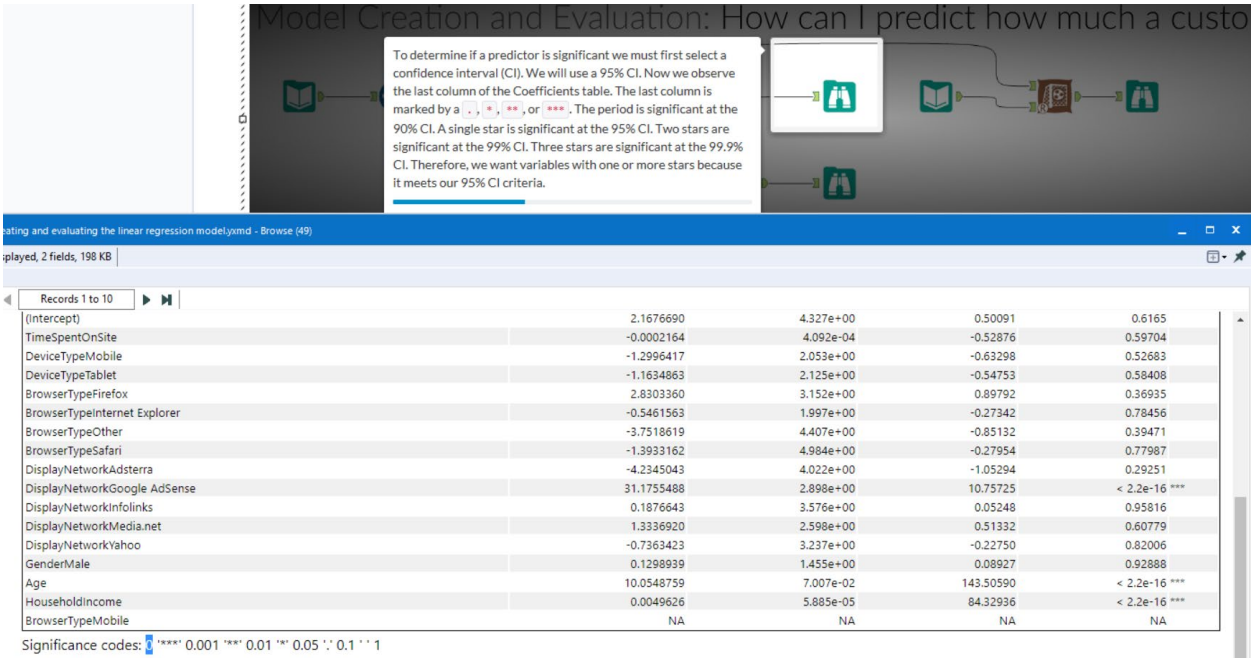
```
IF [Purchased_sale_soda] = "1" THEN IIF([Score_1]>.5,"Correct","Incorrect")
ELSE IIF([Score_1] > .5,"Incorrect","Correct") ENDIF
```

A measure not covered was the model's Akaike information criterion (AIC), which is a measure of the relative quality of statistical models for a given dataset. The AIC by itself does not give a meaningful measure of the model's quality. However, comparing the AIC value for a group of models is a good way to choose a model. A lower AIC value means that a model has some combination of better accuracy and fewer variables. Though the McFadden R-squared value does show which model has better accuracy, it does not penalize models with more variables. The AIC does penalize such models in order to help prevent overfitting. If we were comparing our model to other models, we would probably prefer the model with the lower AIC.

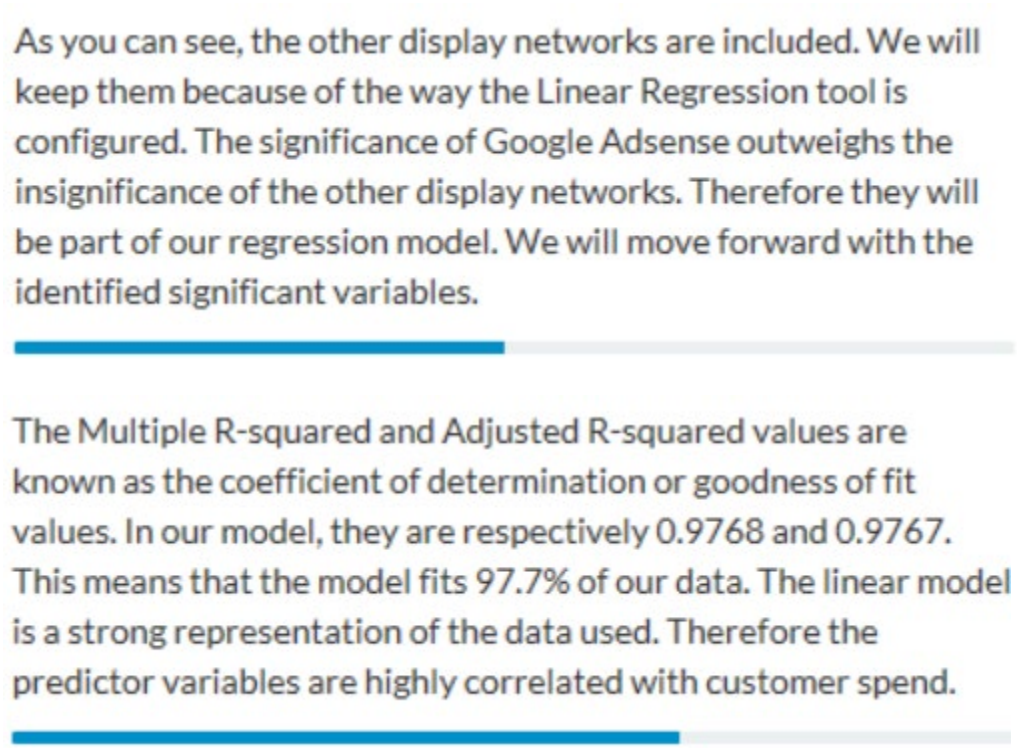
The first step in building a linear regression model is to prepare the input dataset. The Linear Regression tool accepts one input that contains the target and predictor variables.

In order to create a valid linear regression model, your data must satisfy the following conditions:

1. The target variable is quantifiable and continuous — meaning that it can be any value between its minimum and maximum range (e.g. a person's height.)
2. Predictor variables can be continuous or discrete. Examples of continuous variables are temperature, height, and age. Examples of discrete variables are gender, state, and zip code.
3. There is no minimum or maximum number of data points required for linear regressions. However, it is important to balance the number of data points you use to build a model — too few can result in a poor model with a strong bias, too many and you needlessly waste resources.
4. The relationship between target and predictor variables needs to be linear.
5. Other assumptions outside the scope of this kit are:
    - No heteroscedasticity between variables.
    - No multicollinearity between variables.
    - Value of errors are normally distributed.



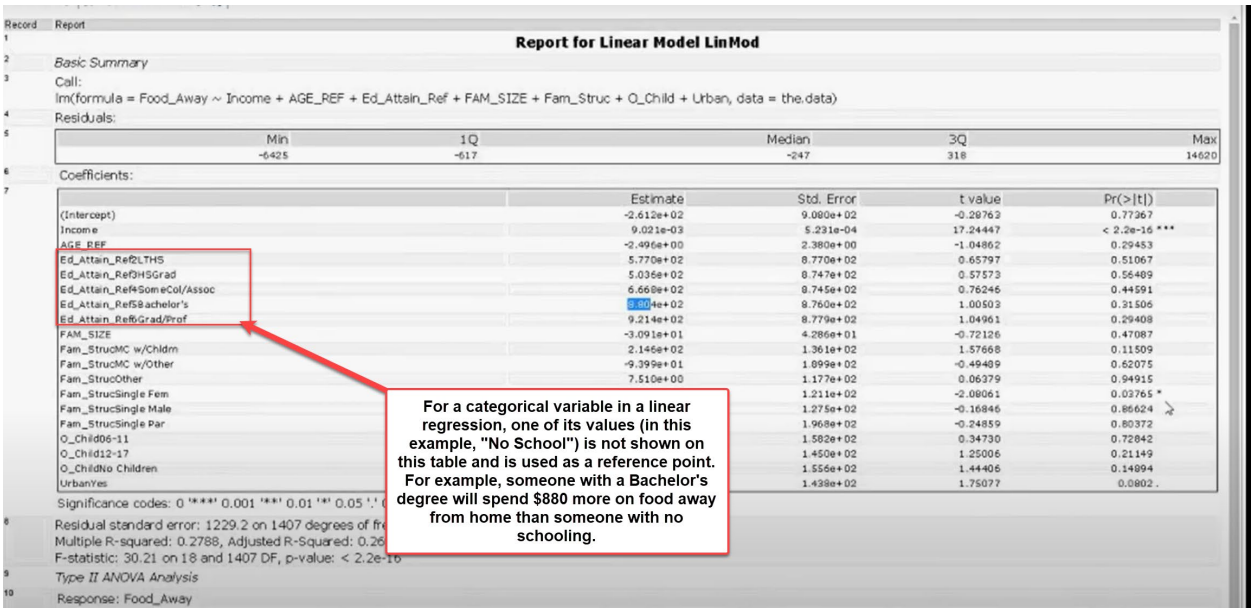
After deselecting fields that do not contribute to the model and rerunning the Linear Regression tool:



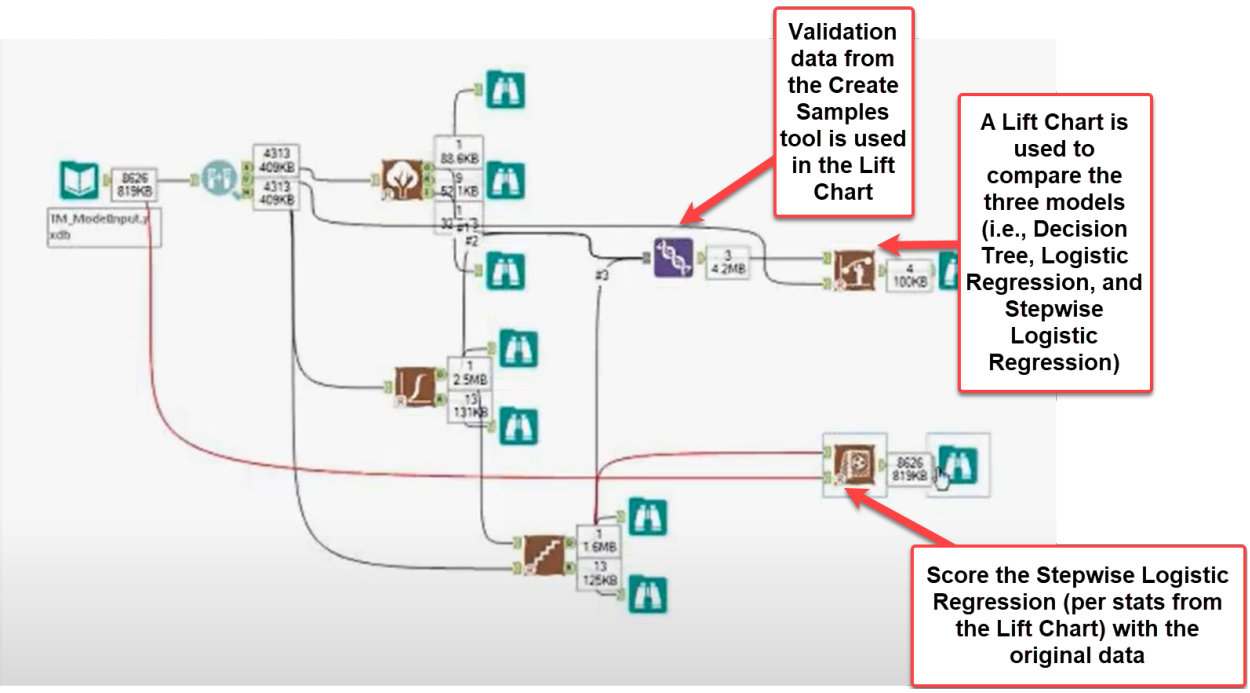
A high R-squared value is a necessary measure of model quality, but is not sufficient when used alone. Be careful when using R-squared. It can be manipulated by adding predictor variables, even if the variables are insignificant. However, the Adjusted R-squared attempts to account for this issue by penalizing models with more variables.

Linear regressions can be utilized across many industries to predict many different target variables. You can predict rainfall, miles driven, weight, or anything else that can be considered a continuous, dependent variable.

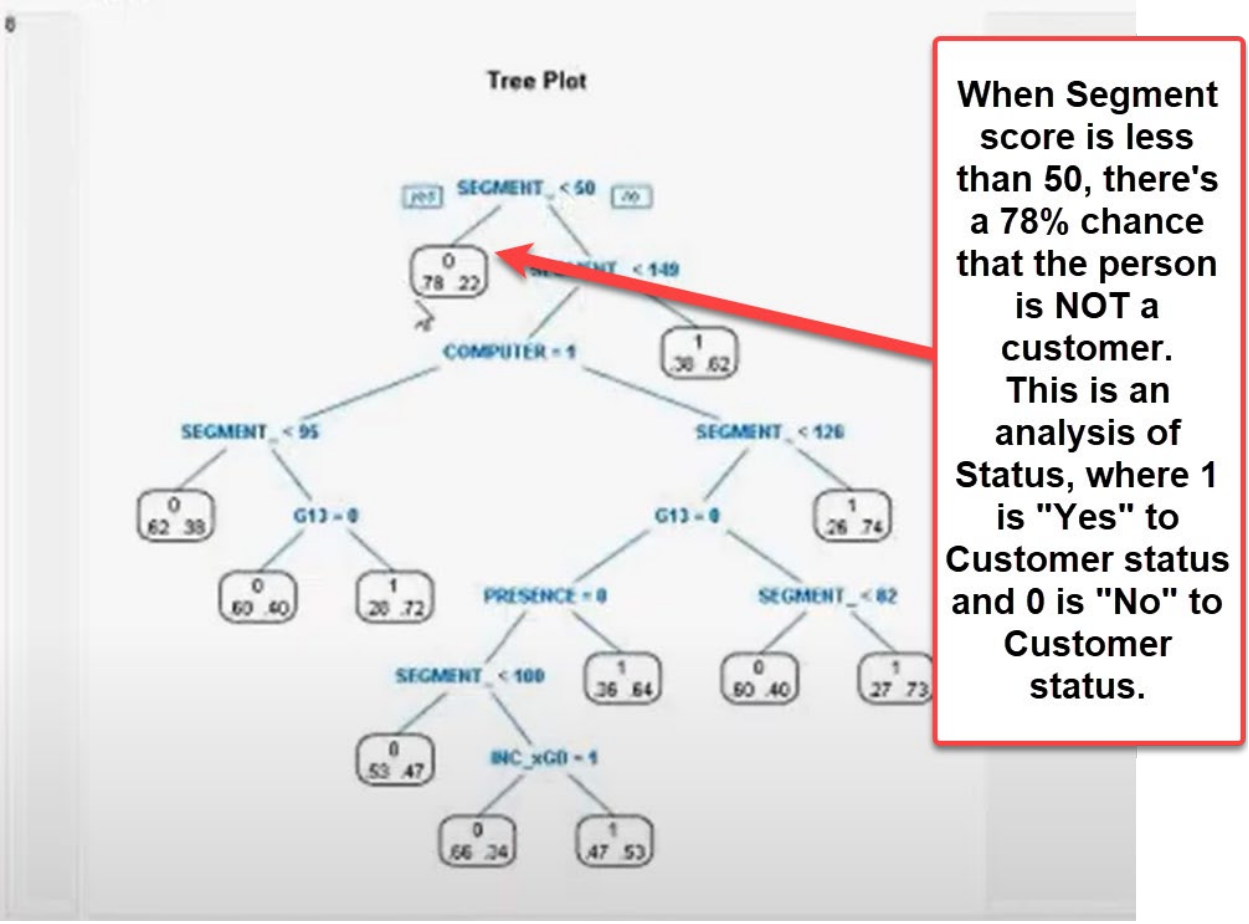
67. Tool Mastery | Association Analysis
68. Tool Mastery | Score
69. Pre-Predictive: Using the Data Investigation Tools - Part 3 of 4
70. Box plot
71. Weekly Challenge #157: An Expert Challenge
72. Linear Regressions and categorical variables



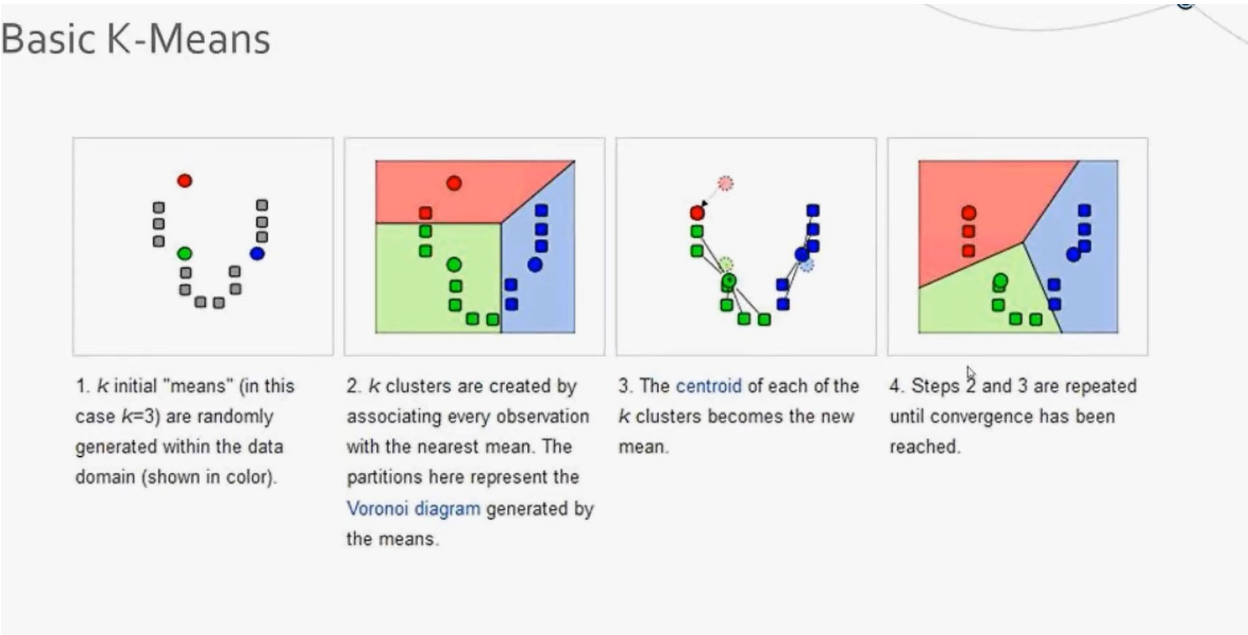
73. Example of a classification problem, from Regression Modeling:



Decision Tree:



74. Basic K-Means:

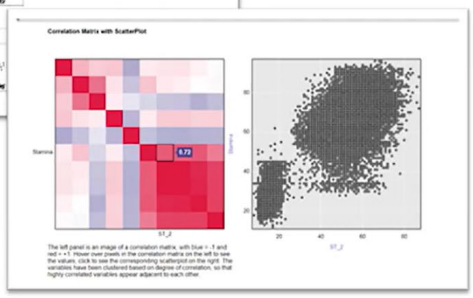


75. [Cluster Analysis Basics]
76. Data Investigation/Prep (Key Terms):

Data Investigation/Prep – Key Terms

- Outlier – a value that is significantly different from other values. These values usually represent an anomaly, a niche case, or an error in the dataset. Defining what constitutes an outlier is discretionary but common measures are >2 standard deviations from the mean or 1.5 times the Inner Quartile Range.
- Collinearity – overlap of predictor variables. Essentially, when two predictor variables are closely related, the amount of new/useful information contained in a row of data used for making the model is decreased.

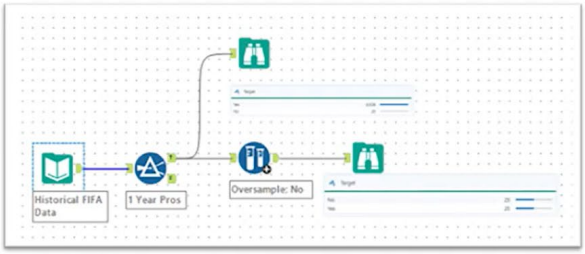
Field Summary - Showing plots of potential outliers



Association Analysis: Showing interactive plots to check of collinearity using Pearson's Correlations

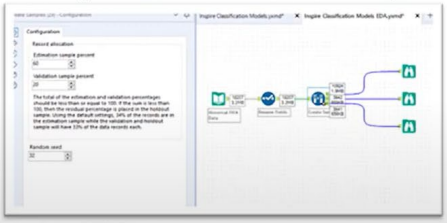
Data Investigation/Prep – Key Terms, Cont.

- Oversampling – a technique used to alter the ratio of the target class (Yes/No) distributions in datasets. This is useful when trying to predict an outcome that is underrepresented in the dataset (e.g., defaults in loan data).
- Creating Samples
  - Used to create Training/Testing/Holdout dataset to develop, evaluate and test models when datasets have not been already separated.
- Other Data Investigation/Prep Considerations
  - Imputation
  - Feature Scaling
  - Binning
  - Encoding
  - Distribution Testing



Oversample:

Showing before and after results of oversample on target field



Create Sample:

Showing before and after results of oversample of target field

77. Logistic Regression characteristics:

Logistic Regression Characteristics

Assumptions

- Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
- Requires the observations to be independent of each other.
- Requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.
- Assumes linearity of independent variables and log odds.
- Typically requires a large sample size

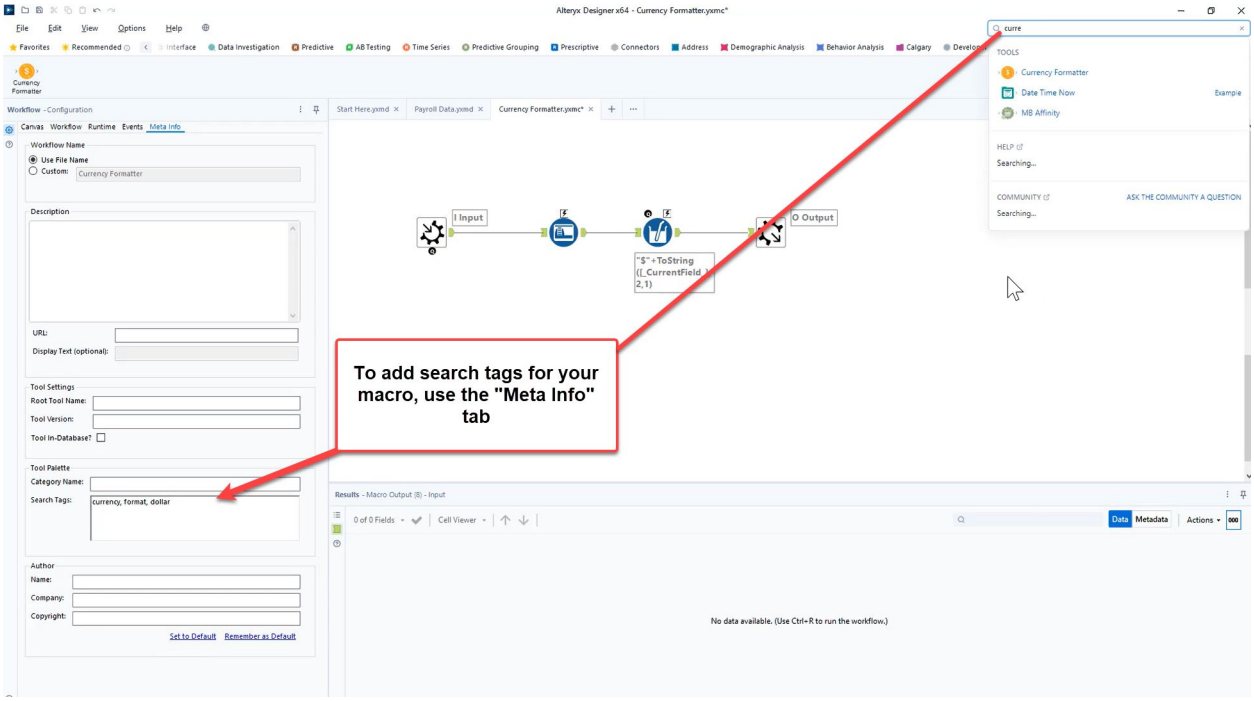
Notes

- Easily impacted by the outliers in the training data
- Performs best when modeling linear associations between classes.
- Pros
  - The linear equation is fairly easy to interpret.
  - Estimation time is relatively short.
- Cons
  - Limited to only binary and ordinal classification.
  - Linear nature of the model has limitations.

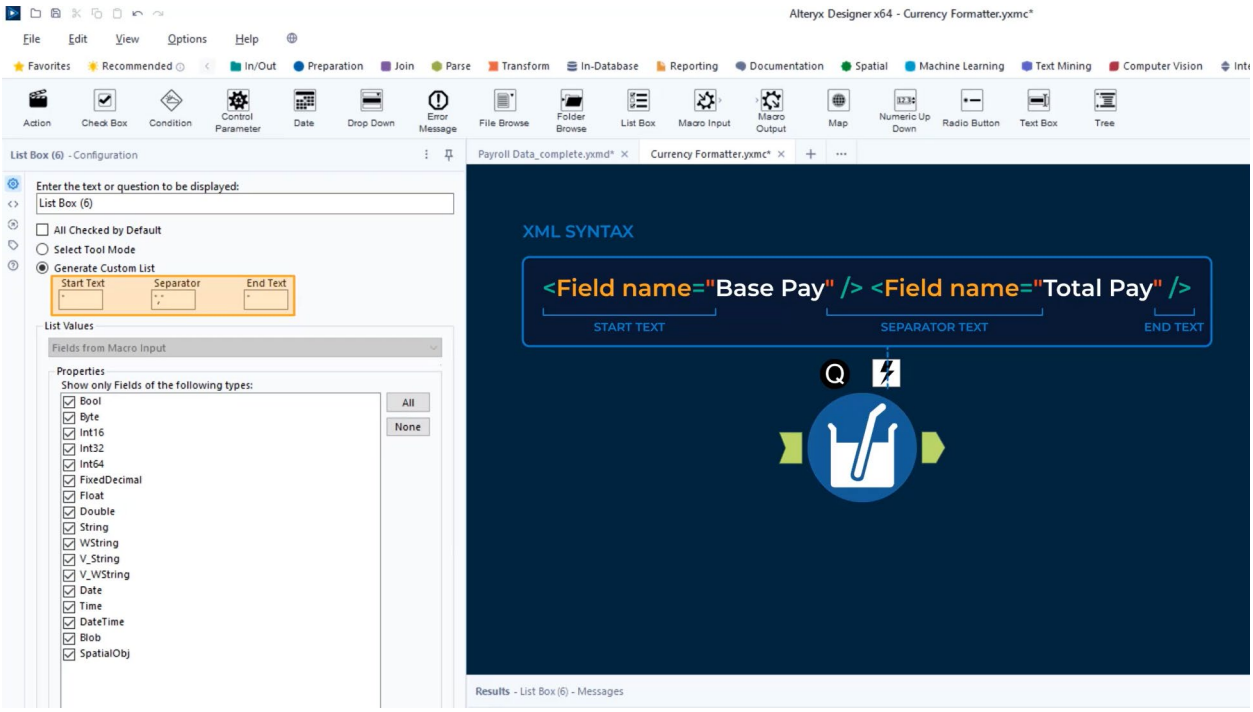
78. Predictive Modeling: Classification
79.

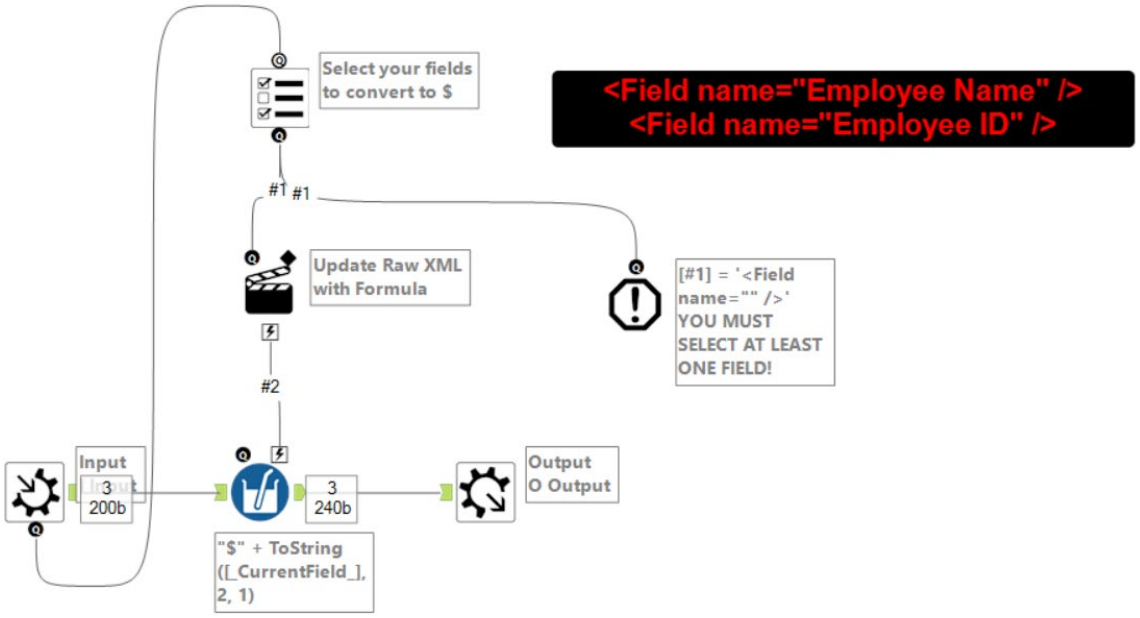
Analytic Applications & Macros

1. Reference shortcuts (and why you should be using them)
2. User constants changed from interface tools
3. Advanced Macro Development
4. Weekly Challenge #39: Trouble Shooting a Broken Macro
5. Weekly Challenge #74: Build a Factorial Calculator
6. Weekly Challenge #118: Think Like a CSE - A not-so-wild-wildcard input!
7. Weekly Challenge #140: Prove the Birthday Paradox!
8. Weekly Challenge #240: Every Vote Counts!
9. Weekly Challenge #246: Rectangle Tangle
10. CS Macro Dev: Conditional Processing with Detours
11. CS Macro Dev: Iterative Macros
    a. The example shown in the link above briefly goes over a very complicated and arduous way to solve the knapsack problem. The article was written before the Optimization tool existed in Alteryx. Nowadays, a much simpler, more efficient approach to solving this problem is by creating a batch macro that loops over the number of boxes to test. It will return the results of, "What is/are the best boxes to take from 1 to n?" This assumes you can use a box once and only once (i.e., you can't use multiple of a box). Creating a batch macro that allows the user to select (via a Check Box tool) whether the upper bound should be infinity (i.e., a single box can be used as many times as the constraint(s) allow) allows us to answer more variations to the knapsack problem.
12. Search tags for a macro can be added within the "Meta Info" tab:

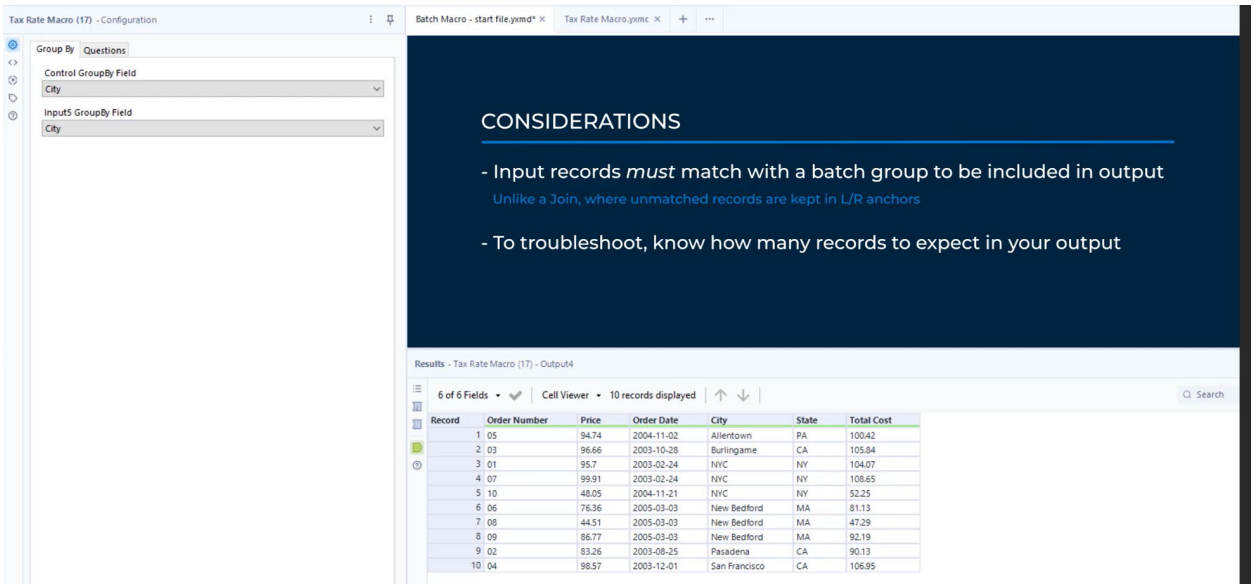


13. Example of editing the XML of a tool by using the selections from a List Box tool:

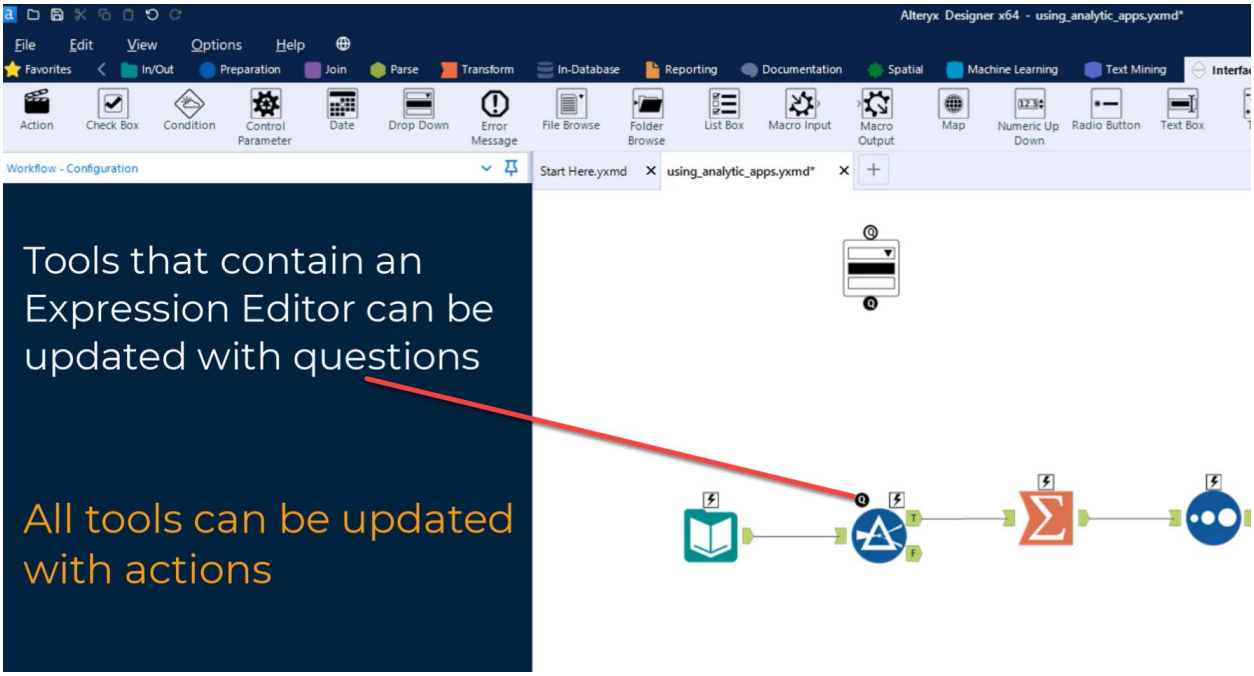




14. [Building dynamic workflows with detour tools or updating XML]
15. Batch macro grouping:



16. [Update Raw XML in Alteryx Macros]
17. [Getting Started with Batch Macros]
18. Tools with an expression editor can be updated with questions:



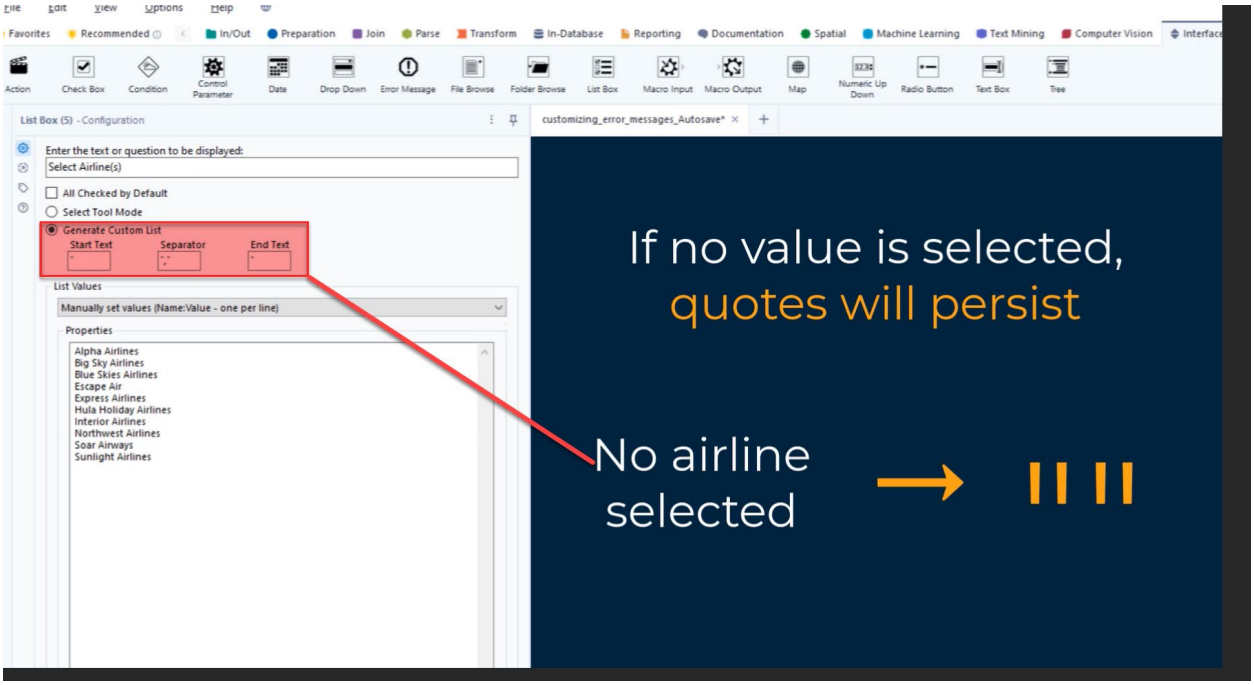
19. List Box questions:

Question 1

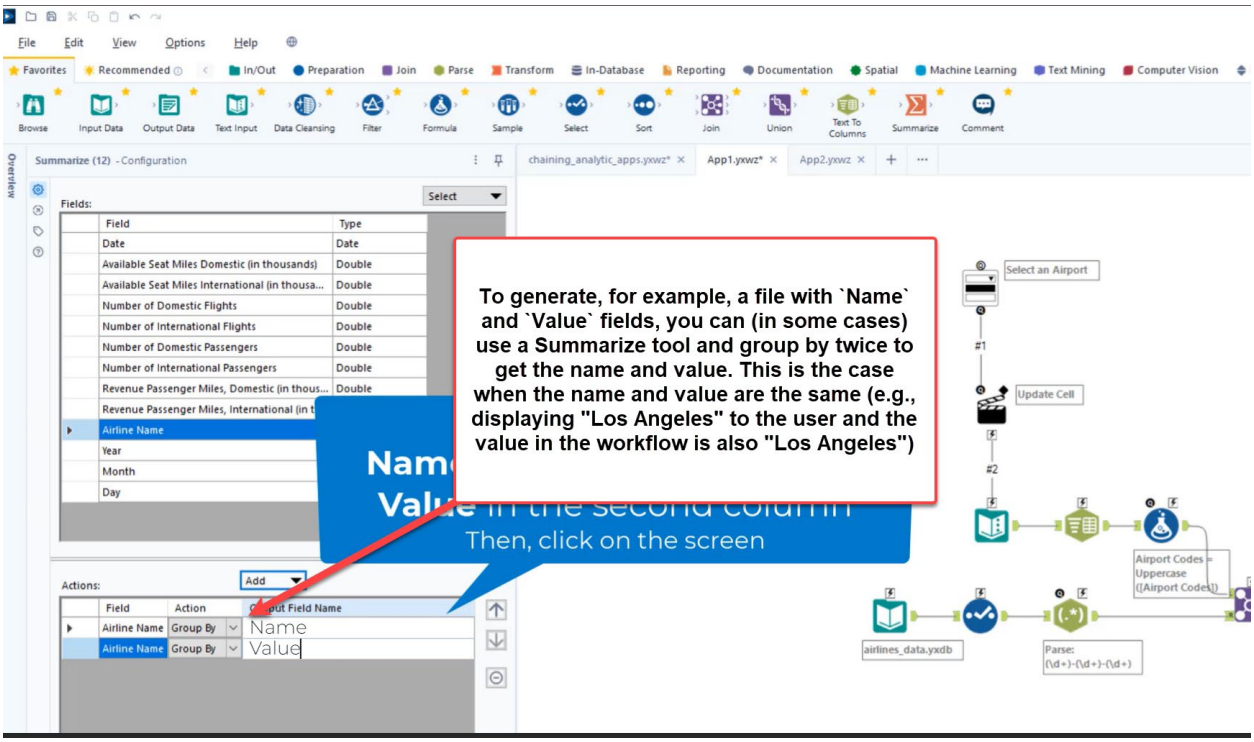
Given the following result from the Debug Workflow's App Wizard, which mode in the List Box was selected in the tool's configuration?

- Select Tool Mode
- Generate Custom List

```
App Values:
<WizardValues>
<Value name="List Box (11)">"JFK","LAX","SFO"</Value>
</WizardValues>
```



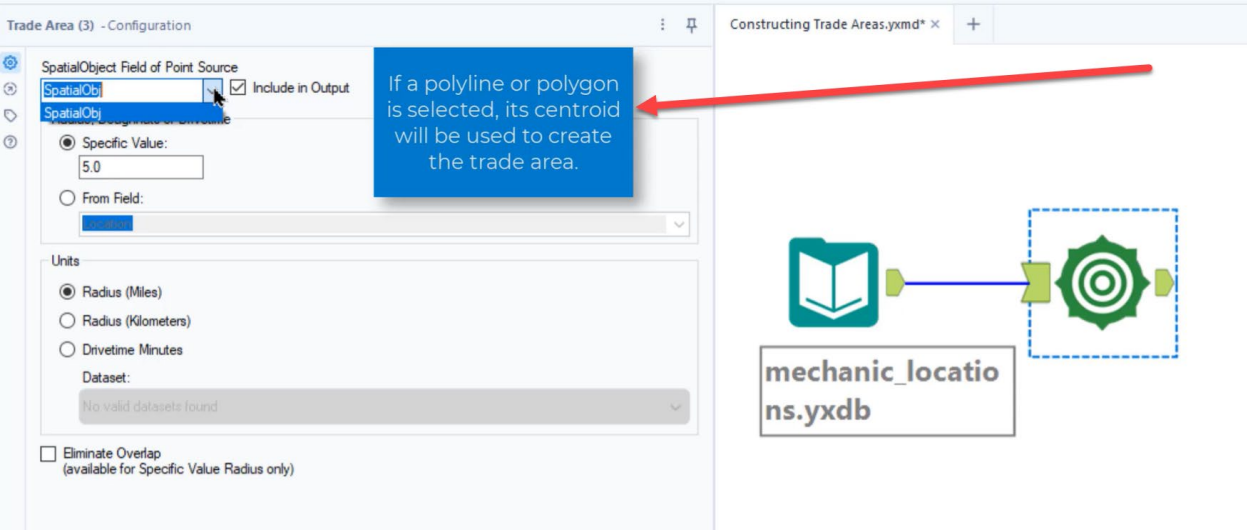
20. Name and value pairs for a chained app may be created using a Summarize tool:



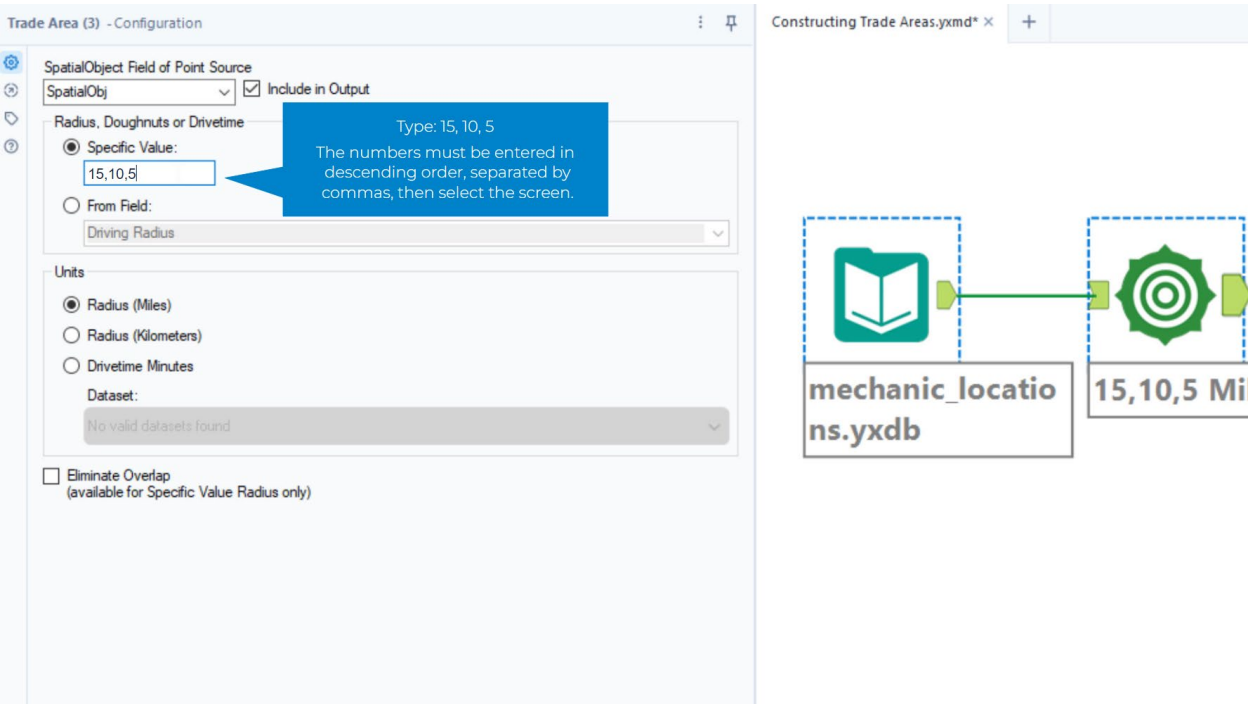
21. [Weekly Challenge #12: Creating an HR Hierarchy]
22. [Building an R Macro]
23. [Using the Data Cleansing Macro]
24. [Santa's First Iterative Macro]
25. [Weekly Challenge #123: When will Rabbits Rule the World?]
26.

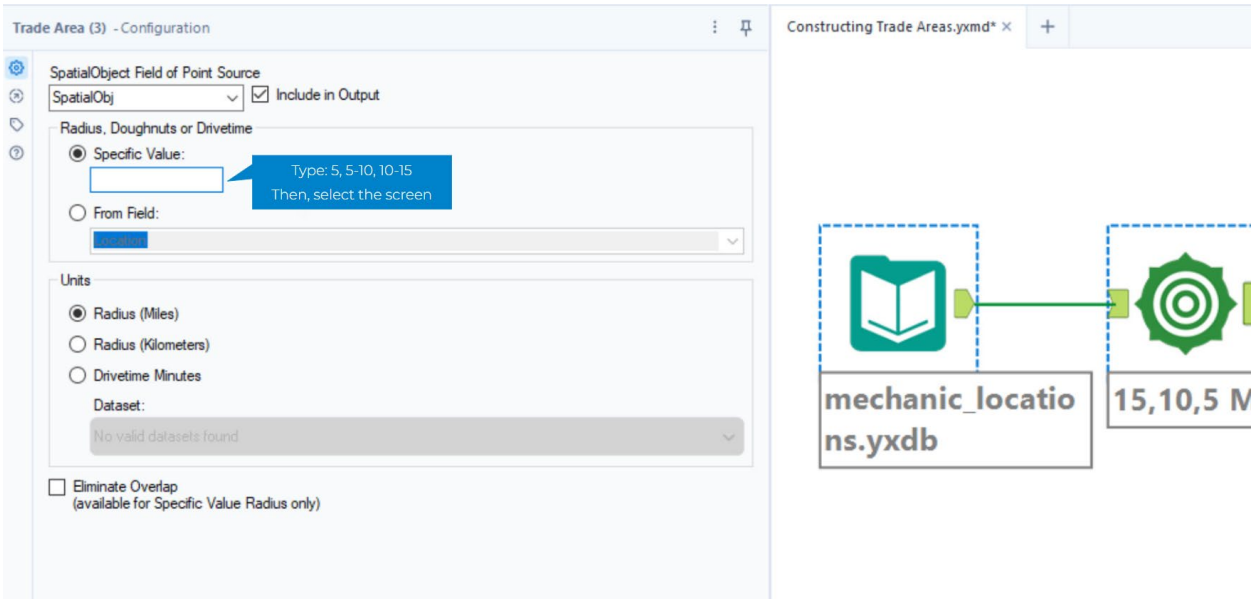
Spatial

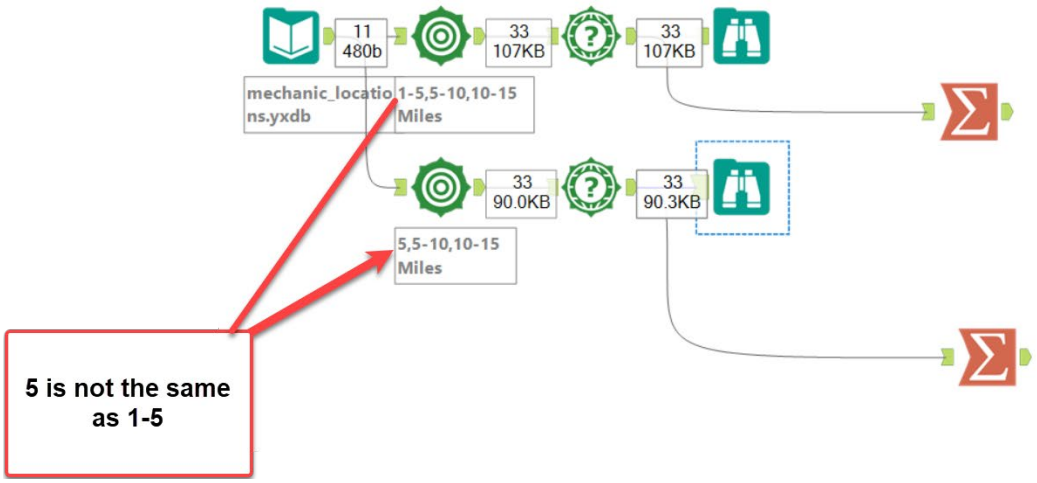
1. [Weekly Challenge #204: Updating Brazil]
2. Trade Area tool:



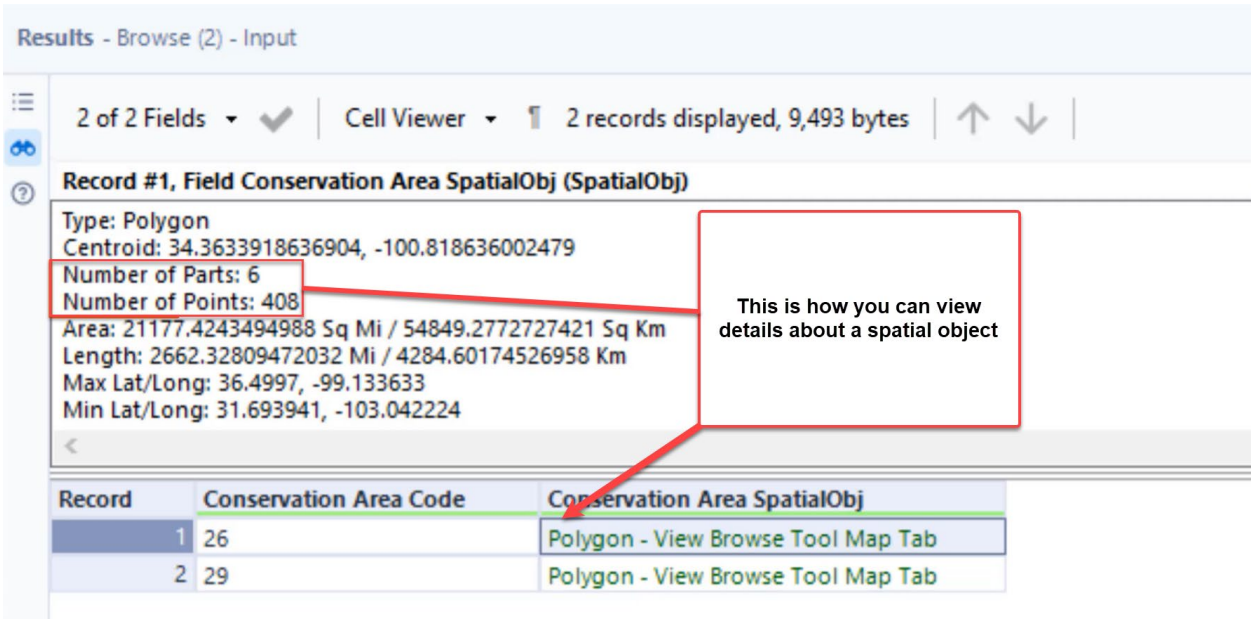
3. Concentric circles in the Trade Area tool: Each larger Trade Area **includes** the smaller Trade Area(s) when using the descending value order, comma separated approach. To create non-overlapping Trade Areas, use a hyphen to separate your Trade Areas (e.g., 5,5-10,10-15). Note that you don't have to follow that specific format. Another example is 1, 3, 5-7.



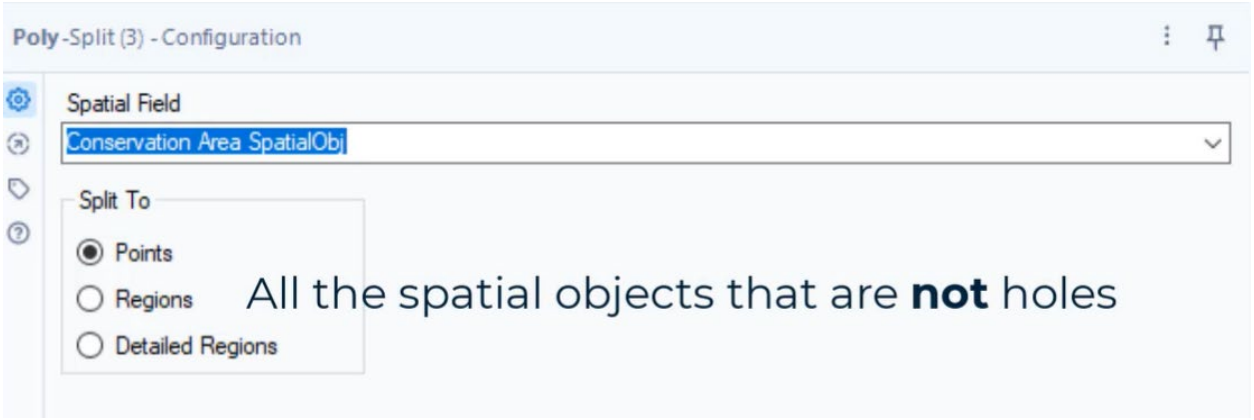


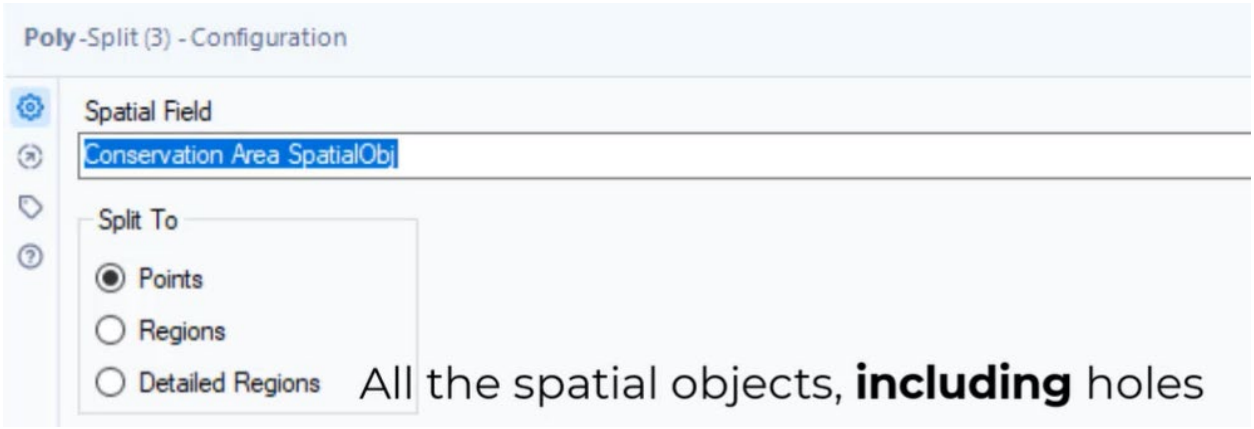


4. Dimension
5. Spatial object details from the Browse tool



6. Poly-Split tool's regions vs. detailed regions

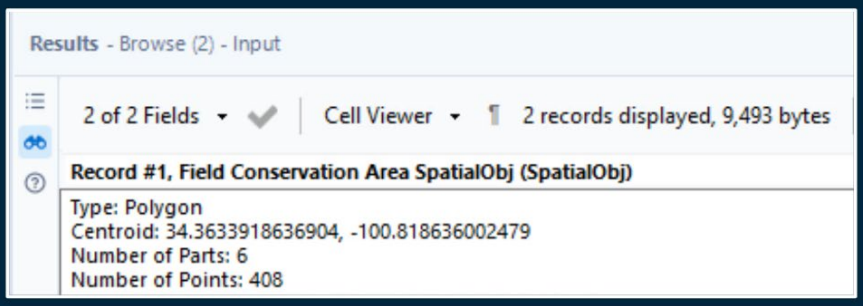




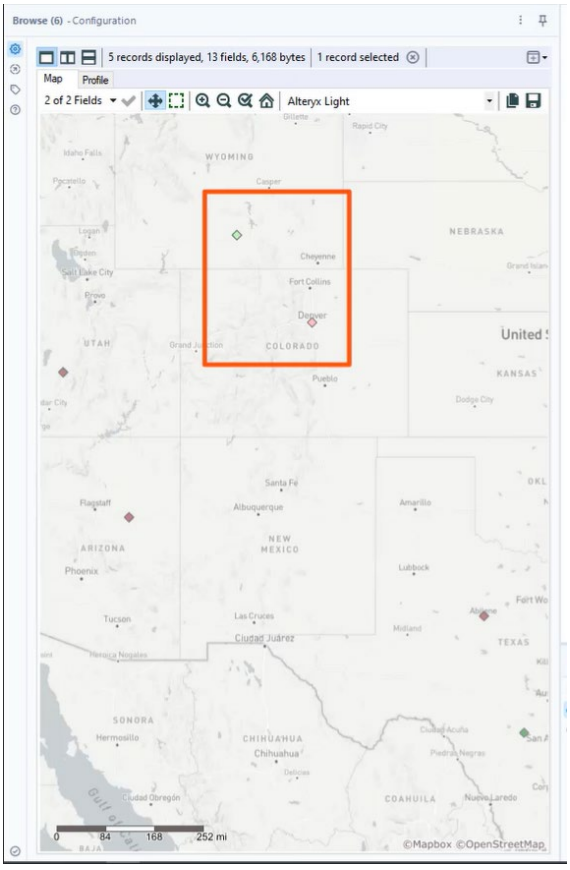
Question 3

Given the information for the following polygon, how many spatial objects could be created if the Poly-Split tool was configured to split the polygon into regions? Select all that apply.

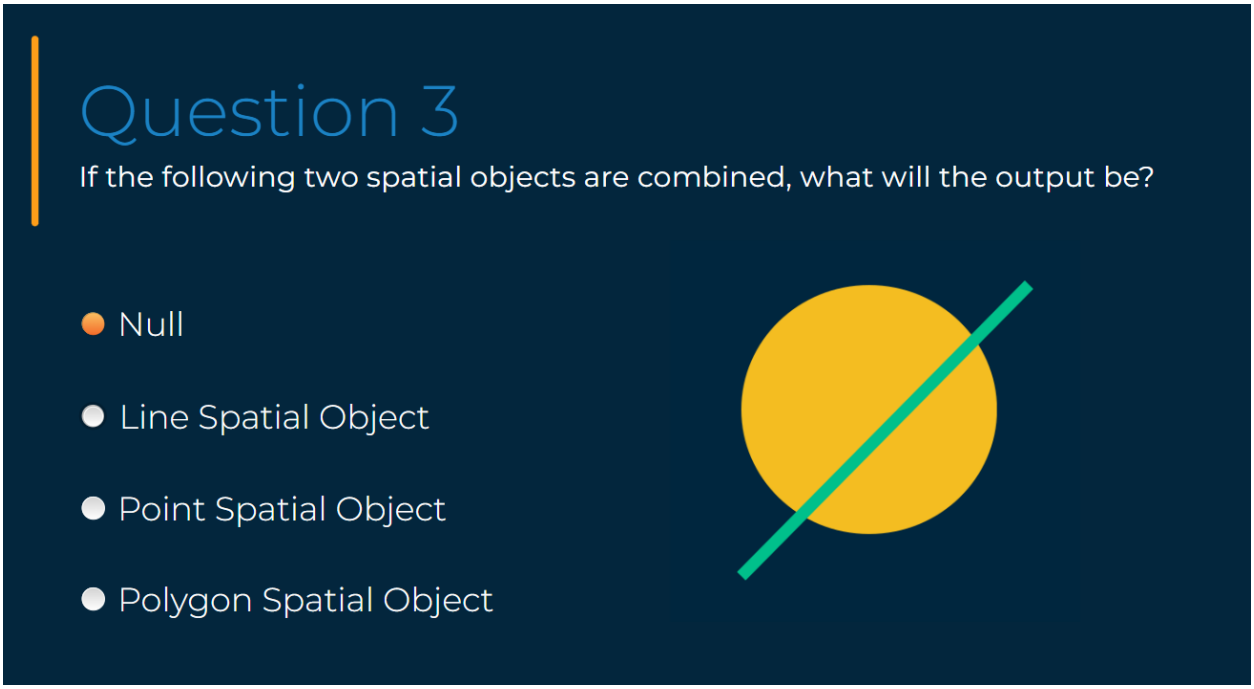
- [ ] 0
- [x] 4
- [ ] 265
- [x] 6



7. In the Browse tool, the source used in a Distance tool is shown in pink and the destination is shown in green.



8. Only spatial objects of the same type, such as two polygons, can be combined. Combining a point and a line will produce a [Null]. If you use have two different types of spatial objects (e.g., a line and a polygon) that intersect, the simpler spatial object (in this example, a line) will be returned.



9. Spatial relationships

Spatial Relationships

The Spatial Match tool relates a Target spatial object to a Universe spatial object based on a specified relationship. There are 4 basic types of spatial relationships: touch, contain, within, intersect.

Touch

A Target object touches a Universe spatial object if the Target touches the outside boundary of the Universe object but does not share any of the same interior space.

Contain

A Target object contains a Universe object if the Target object completely encompasses the Universe object.

Within

A Target object is within a Universe Object if it is completely encompassed by the Universe object.

Intersect

A Target object intersects a Universe object if the two overlap, or share any area in common.

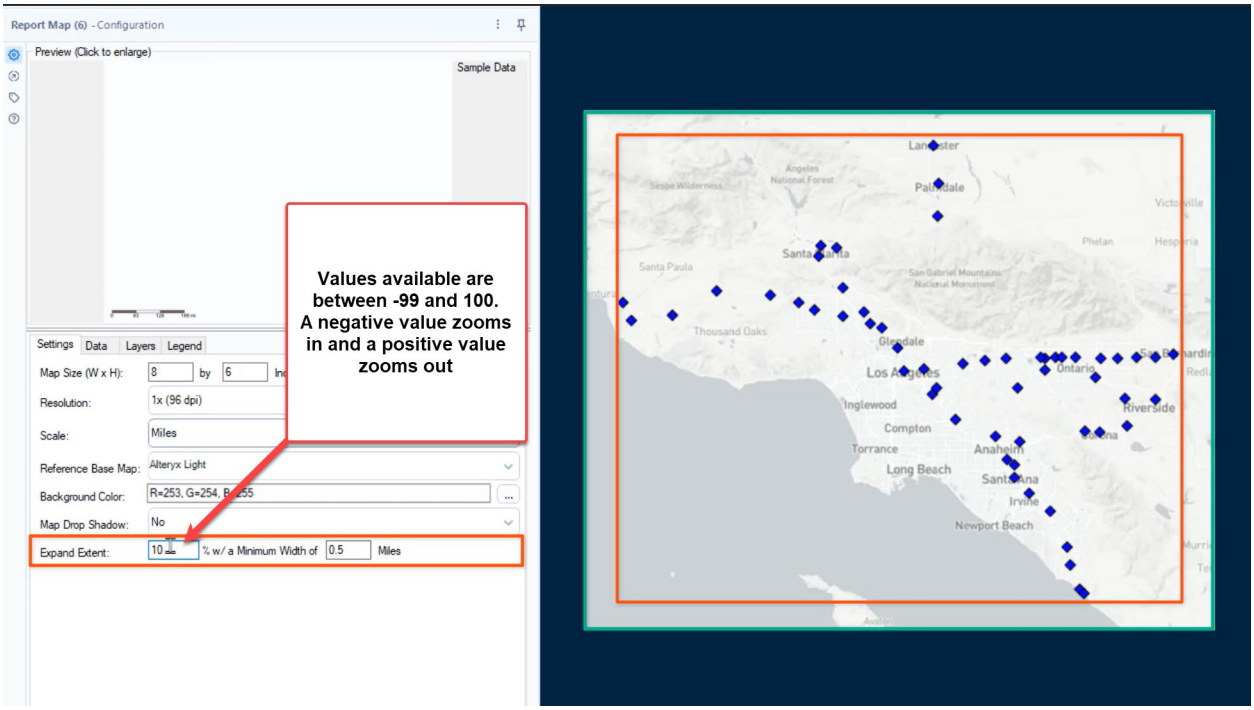
10. Intermediate Spatial Analytics
11.

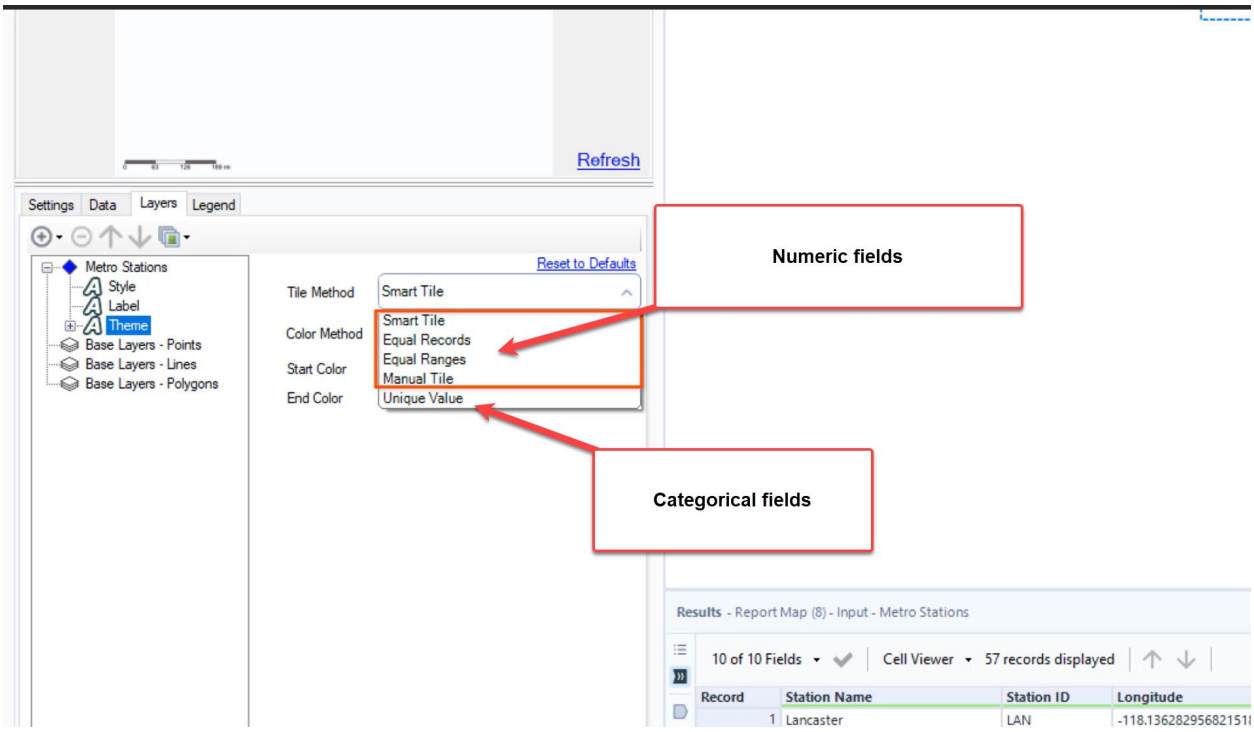
Reporting

1. An example of using a group by within the Table tool

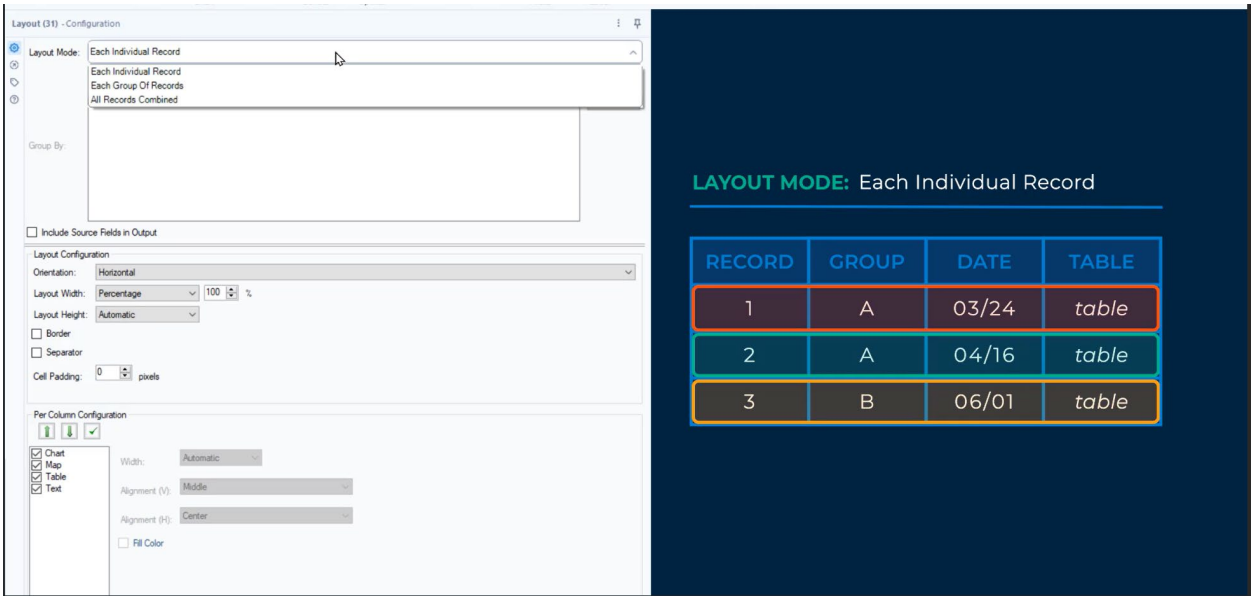
Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Palmdale	Antelope Valley	1,829	2,180	$8
West Corona	Inland Empire	788	1,015	$13
Union Station	Los Angeles	13,645	16,097	$14
Commerce	Orange County	1,982	2,096	$17
Riverside-Downtown	Riverside	2,283	1,496	$26
Montclair	San Bernadino	2,392	3,233	$28
Montalvo	Ventura County	1,169	1,330	$15

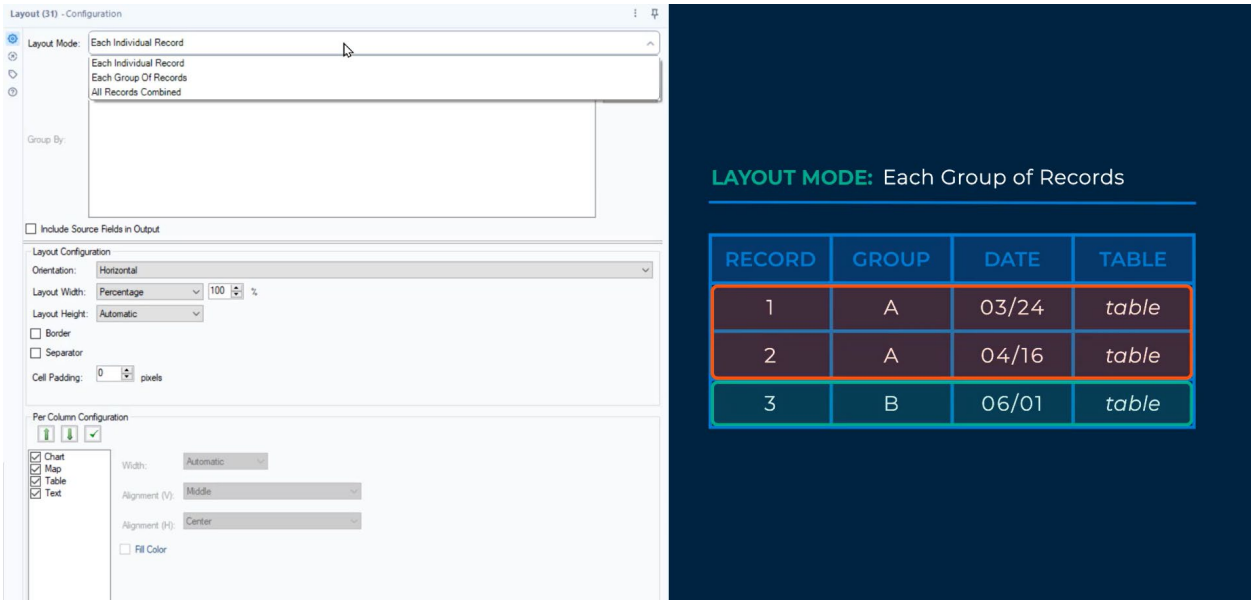
2. Report Map

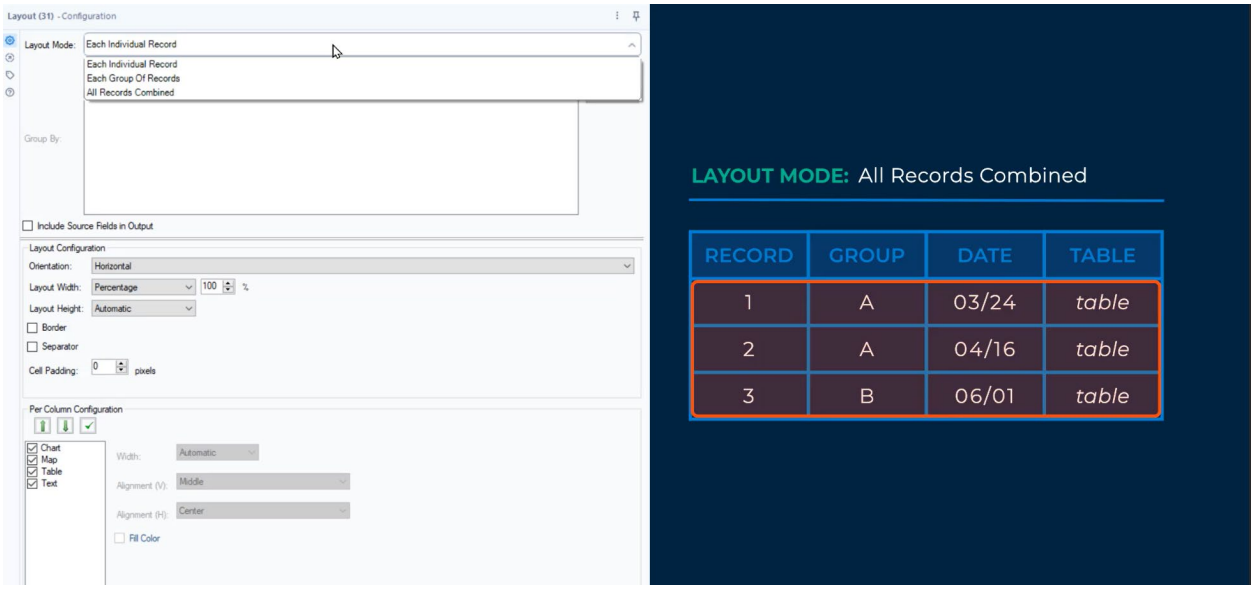



Refresh
Settings
Data
Layers
Legend
Reset to Defaults
Metro Stations
Style
Label
Theme
Base Layers - Points
Base Layers - Lines
Base Layers - Polygons
Tile Method
Smart Tile
Color Method
Equal Records
Equal Ranges
Manual Tile
Start Color
End Color
Unique Value
Numeric fields
Categorical fields
Results - Report Map (8) - Input - Metro Stations
10 of 10 Fields
Cell Viewer
57 records displayed
Record
Station Name
Station ID
Longitude
1
Lancaster
LAN

3. Layout


Layout (31) - Configuration
Layout Mode:
Each Individual Record
Each Individual Record
Each Group Of Records
All Records Combined
Group By:
Include Source Fields in Output
Layout Configuration
Orientation:
Horizontal
Layout Width:
Percentage
Layout Height:
Automatic
Border
Separator
Cell Padding:
pixels
Per Column Configuration
Chart
Map
Table
Text
Width:
Alignment (V):
Alignment (H):
Fill Color
LAYOUT MODE: Each Individual Record
RECORD
GROUP
DATE
TABLE
1
A
03/24
table
2
A
04/16
table
3
B
06/01
table

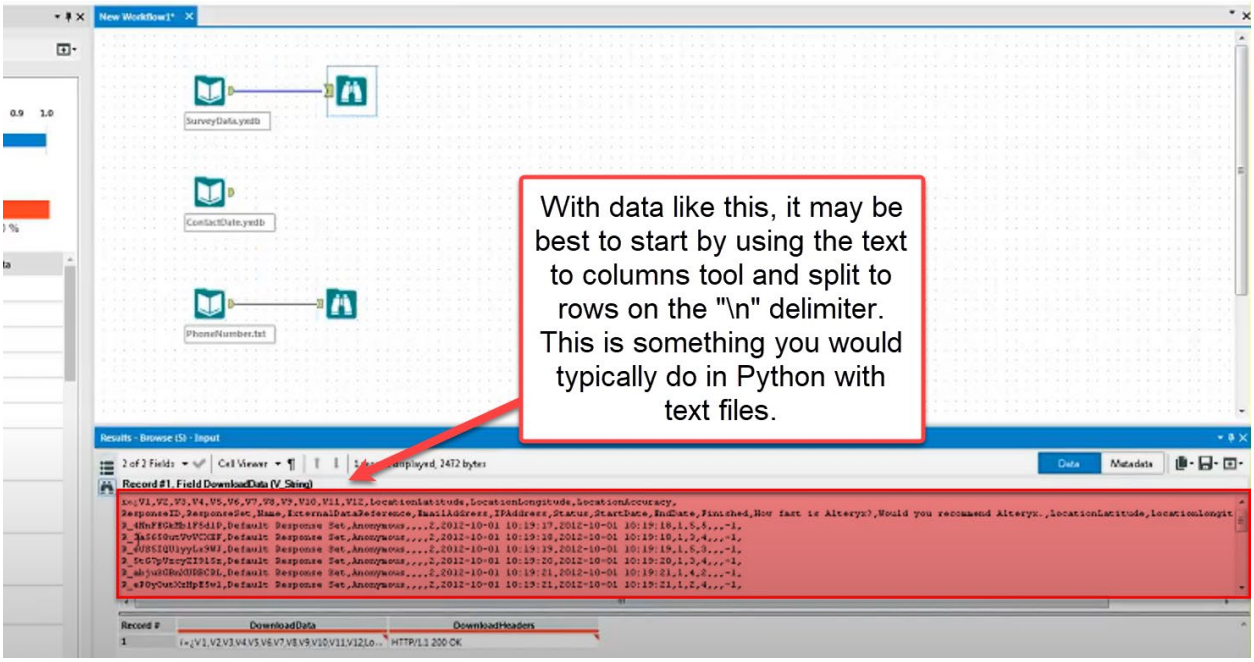




4. [Getting Started with Batch Reporting]
5. [Getting Visual with Reporting]
6. [Batch Reporting: Building Tailored Reporting Processes]
7.

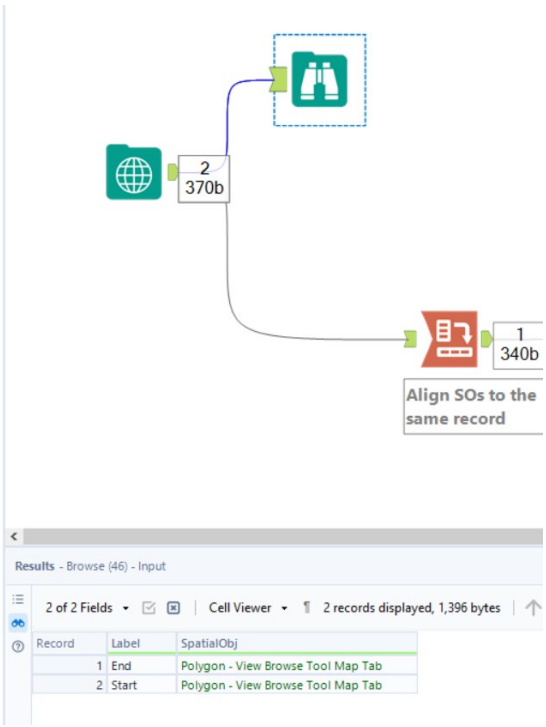
In-Database and General

1. How do Alteryx's In-Database tools work?
2. In-Database Workflows
3. New line (i.e., \n) delimited data:

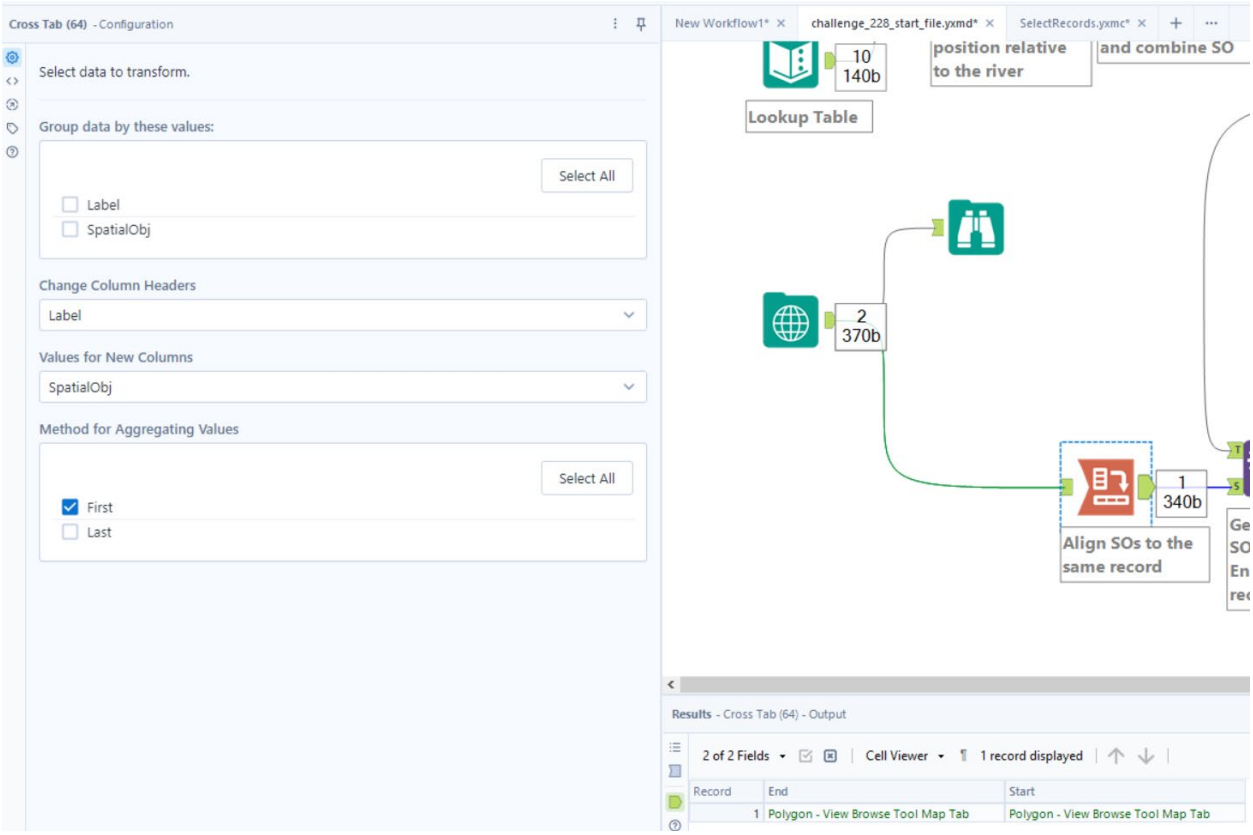


4. Remember to use the Cross Tab tool instead of, for example, the Sample tool to align to the same record.

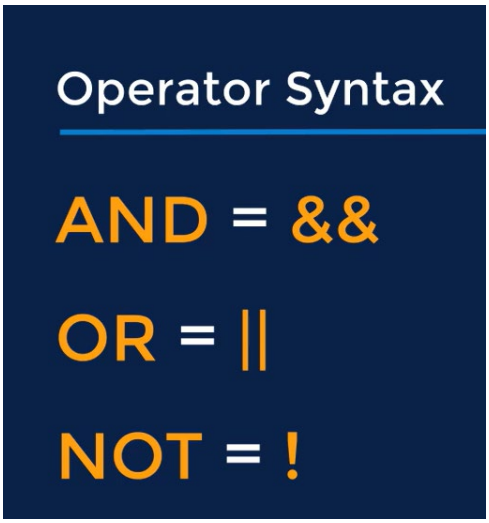
Before:



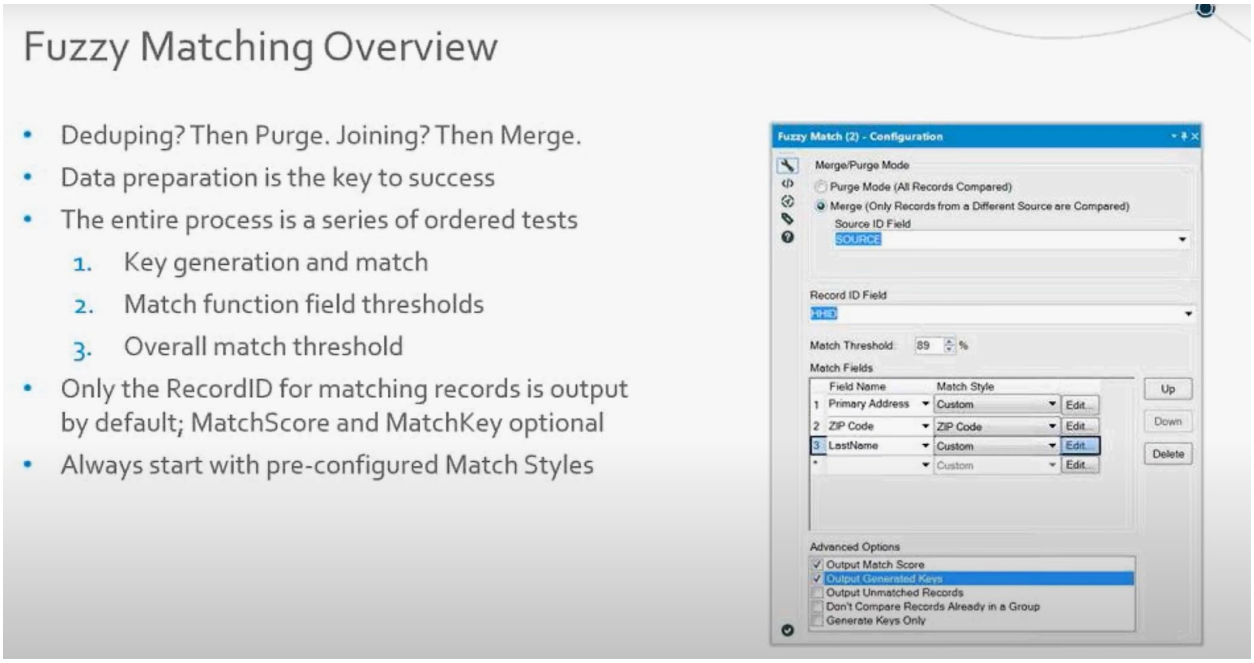
After:



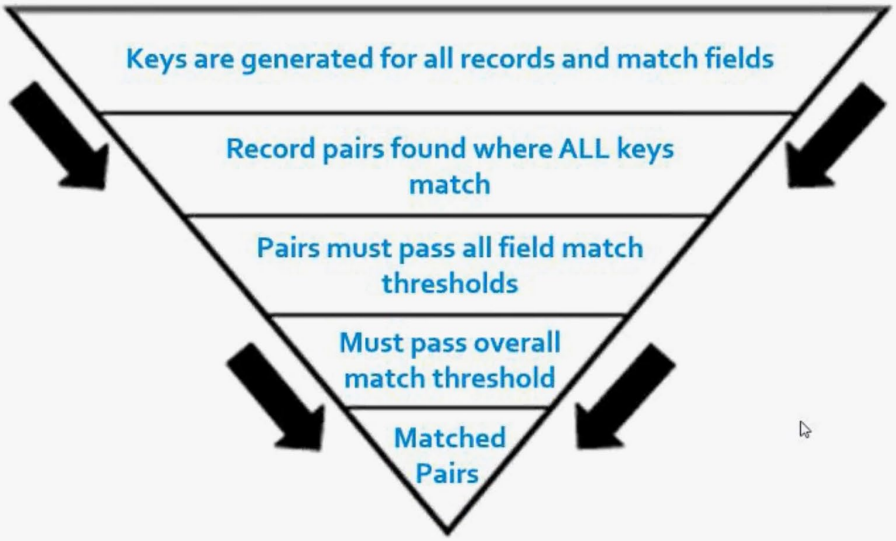
5. Operator syntax



6. Dynamic Input training
7. Running Total and Weighted Average
8. Parsing XML
9. Fuzzy Matching



The Fuzzy Match Process: Behind the Scenes



Key Generation

Tips & Tricks:

- Key generation is used to narrow down the population of possible matches. Keys can be generated for all fields or some fields, but at least one field must have a key generated.
- Can choose "no key generation" strategically when using multiple fields. For instance, when trying to match with nicknames, key generation isn't advised.
- Shorter key length can result in more potential matches.

First Name	Last Name	MatchKey
CAILIN	SPINGLE	KLN|SPNK
KAITLIN	SWINGLE	KTLN|SNKL
KAITLIN	SWINGLE	KTLN|XNKL
KAYLEN	SWINGLE	KLN|SNKL
DAVID	SWINGLE	TFT|SNKL
DAVE	SWINGLE	TF|SNKL
DAVE JR.	SWINGLE	TFT|SNKL

10. Fuzzy Matching - Intermediate Users
11. The Art and Science of Fuzzy Matching
12.

In-Database Analysis

- Create new in-database connection strings
- Connect to SQL database tables
- Stream data into an in-database workflow
- Analyze data using in-database tools and appropriate SQL syntax
- Change data types and sort data in-database
- Create and update tables in the database

Predictive Analysis

• Investigate and prepare data for analysis

• Identify suitable variables for predicting a target variable

• Select predictor variables that fit given criteria

• Create training and validation datasets

• Select appropriate algorithms to model datasets

• Train models

• Compare models

• Interpret model reports

• Score new data with trained models

• Classification

• Regression

• Time series

• Optimization

Spatial Analysis

- Construct and manipulate spatial objects
- Establish and quantify spatial relationships between objects
- Use spatial relationships to create new spatial objects
- Conduct complex, multistep spatial area analysis
- Leverage spatial data to determine optimal routing and delivery paths
- Evaluate spatial relationships to calculate spatial coverages

Analytic Apps & Macros

- Create complex batch and iterative macros
- Create chained applications
- Create complex, dynamic applications with cascading logic and conditional functionality
- Troubleshoot and optimize provided macros
- Create an analytic application from a provided workflow

Reporting

- Create and output dynamic, batched reports

with the following elements:

- o Headers and footers
- o Tables and charts
- o Images and text
- o Specific layout criteria
- o Report maps