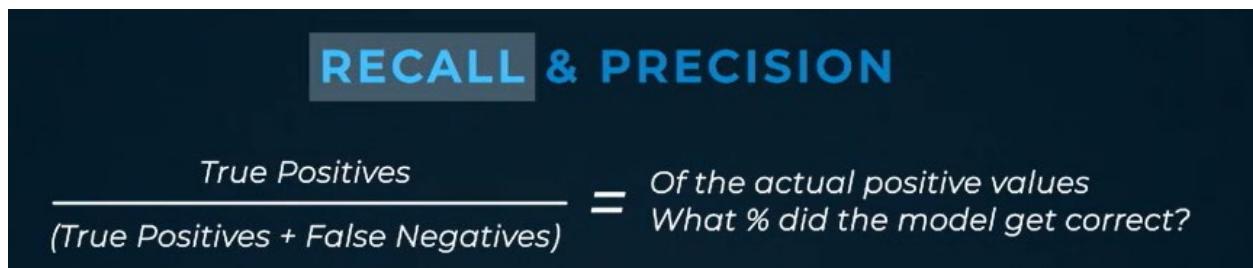
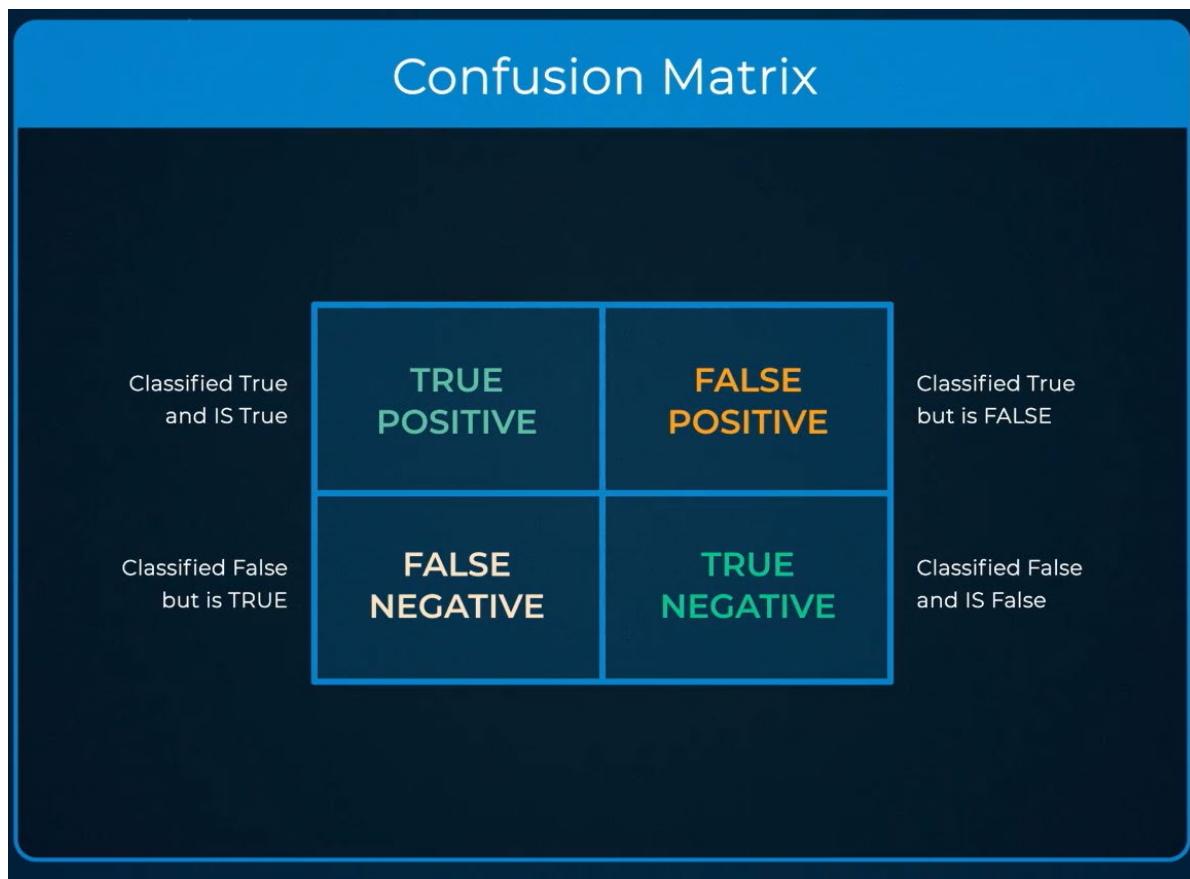


Expert Exam

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

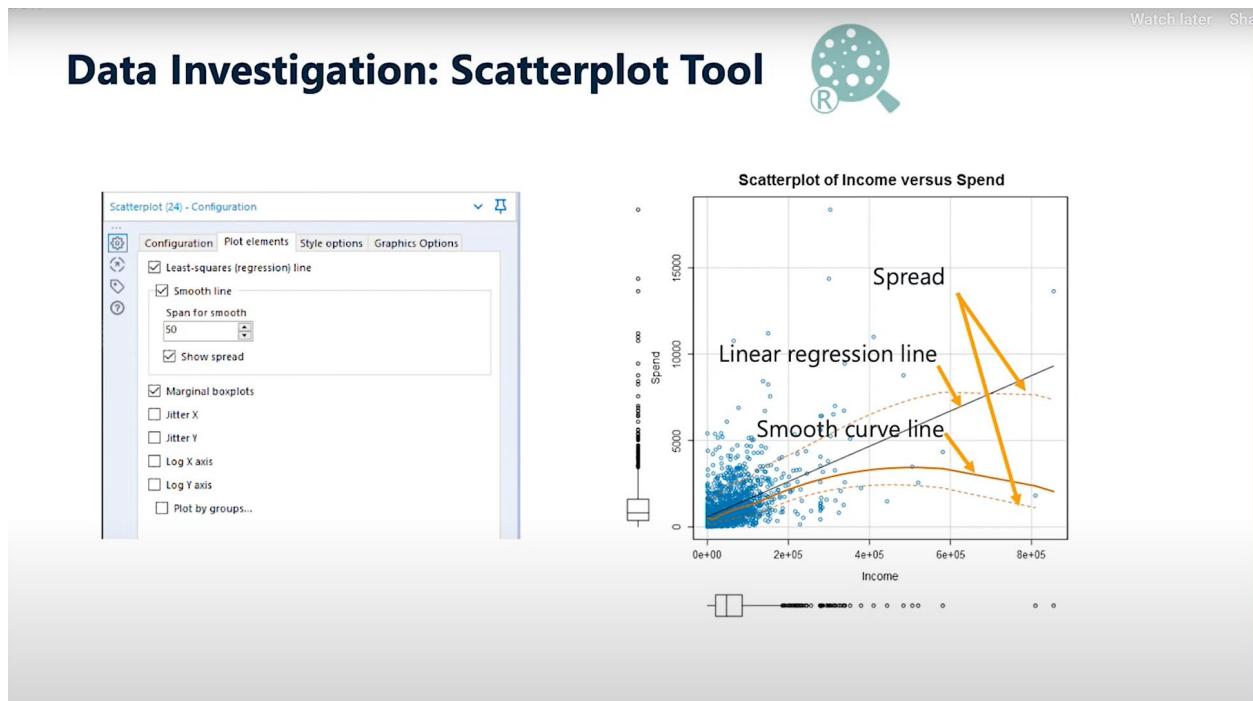
1. [How to Interpret Pr\(>|t|\) in Regression Model Output in R](#)
2. [How to Interpret Significance Codes in R](#)
3. [How to Calculate a P-Value from a T-Test By Hand](#)
4. [Live Training: Regression Modeling](#)
5. [Mastering Time Series Analysis & Forecasting in Alteryx](#)
6. [Confusion Matrix](#)



RECALL & PRECISION

$$\frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})} = \begin{array}{l} \text{Of the values classified as positive} \\ \text{What \% did the model get correct?} \end{array}$$

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



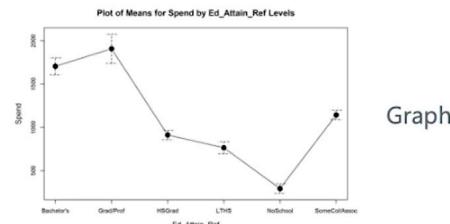
16. Plot of Means

Data Investigation: Plot of Means Tool

One Output

Compare two variables:

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



Graph

17. Action

Action

Watch later S

- 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*.

DATA CLEANSING

Ensure accuracy & remove outliers

Categorize variables as numeric or categorical

Remove constants

Check Datatypes – **Zero = 0**

Look for biases in data collection

Normalize (center) data

Decide how to treat missing values

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

MODELING TOOLS

CLASSIFIERS



Assign a label (sort)

Discrete variable (can be counted)

- Binary: 1 or 0, True or False, Yes or No
- Multinomial: 3 or more groups



REGRESSORS



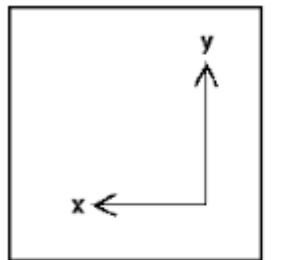
Find a specific value

Continuous variable

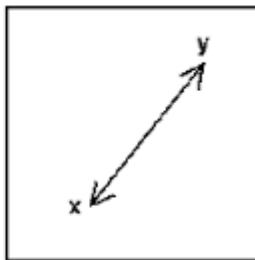
- Usually non-integer



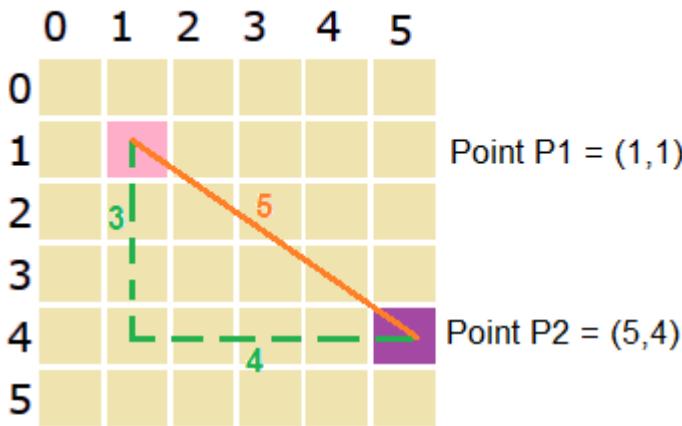
21. Manhattan vs Euclidean (aka “as the crow flies”):



Manhattan



Euclidean



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

$$\text{Manhattan distance} = |5-1| + |4-1| = 7$$

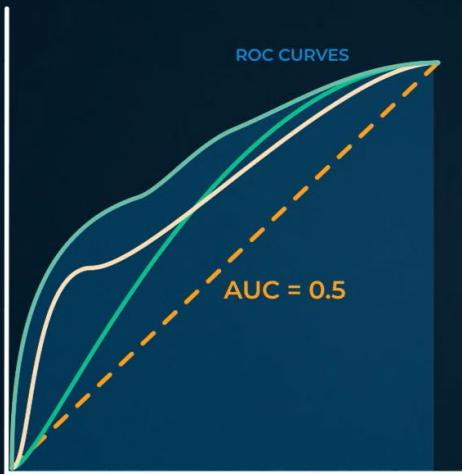
22. Collinearity

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)



Area Under Curve (AUC)

- Total possible area is **1.0**
- Higher AUC is better
- Models that are up and to the left have higher AUC and are usually considered better performers

Question 5

Which ROC Curve shows the best performing model?

- 
- 
- 

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

Question 4

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

- 51
- 55
- 60
- 25

Confusion Matrix		
	Yes	No
Yes	25	30
No	26	35

25. Predictive Analytics question

Question 6

Which of these is the function for accuracy?

- $\frac{TP+FP}{TP+FP+FN+TN}$
- $\frac{TP}{TP+FN}$
- $\frac{TP}{TP+FP}$
- None of these

26. Data Investigation courses of action



27. Bias & Skew

BIAS & SKEW

Occurs when sample does not accurately represent a population

Be mindful of where and how your data was collected

Is your sample data actually indicative of your population?

Investigate your outliers

Unconscious bias is common

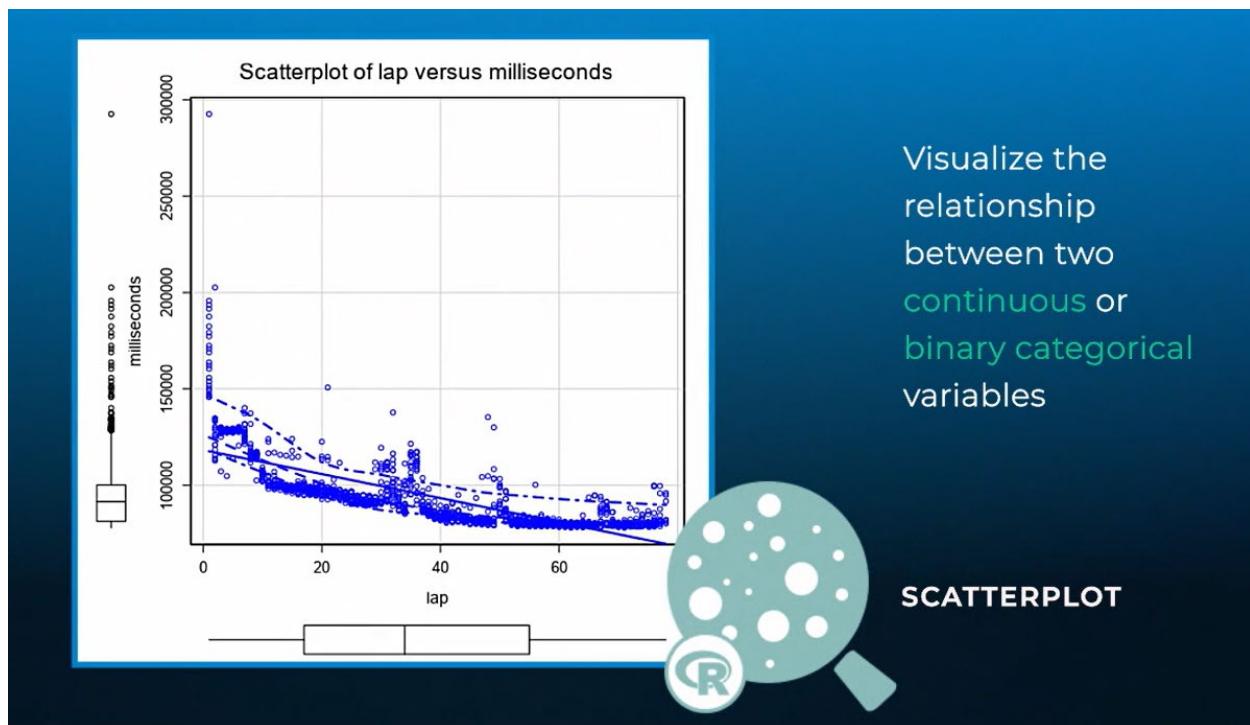
Consider the legal and ethical ramifications of your model

28. Frequency Table

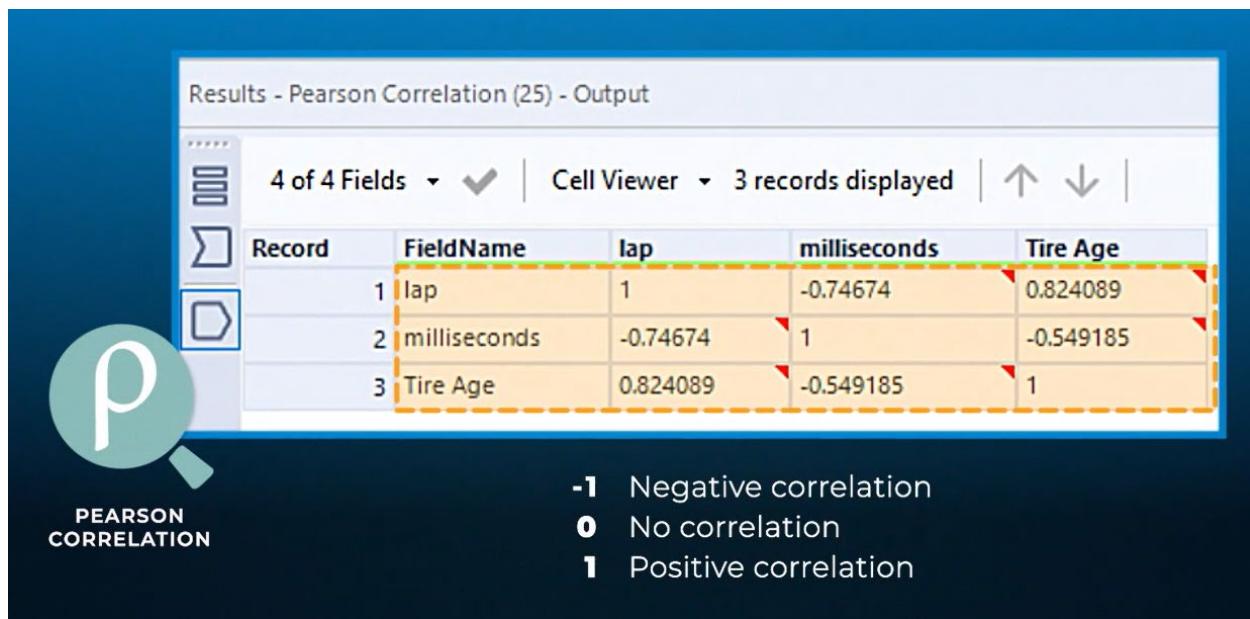
The screenshot shows a software interface for viewing frequency data. On the left, there is a sidebar titled "Data Types Not Accepted:" which lists several data types that are not supported: Fixed Decimal, Float, Double, Date/Time, Blob, and Spatial Object. The main area is titled "FREQUENCY TABLE" and contains a table with the following data:

Field_Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
U	336	24.76	336	24.76
W	336	24.76	672	49.52
I	288	21.22	960	70.74
SS	251	18.50	1211	89.24
S	145	10.69	1356	99.93
[Null]	1	0.07	1357	100.00

29. Scatterplot



30. Pearson Correlation. Only works for linear relationships.



31. Spearman Correlation.

Results - Spearman Correlation (26) - Output16

1 of 1 Fields | Cell Viewer | 1 record displayed

Record	Result
1	-0.697835

Evaluates the monotonic relationship between two variables

 SPEARMAN CORRELATION

Results - Spearman Correlation (26) - Output16

1 of 1 Fields | Cell Viewer | 1 record displayed

Record	Result
1	-0.697835

 SPEARMAN CORRELATION

- 1 Always trends negatively
- 0 No correlation
- 1 Always trends positively

Variables must be **continuous** or **ordinal**

Ordinal variables must be **ranked**

<- Can't say

that *red* is a step up from *green*, but *large* IS greater than *small*.

Spearman correlations indicate a consistent directionality between two numeric variables.

32. Data Investigation question.

Question 1

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

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

Size	Length
Green	12.2
Purple	15.4
Gold	19.6

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

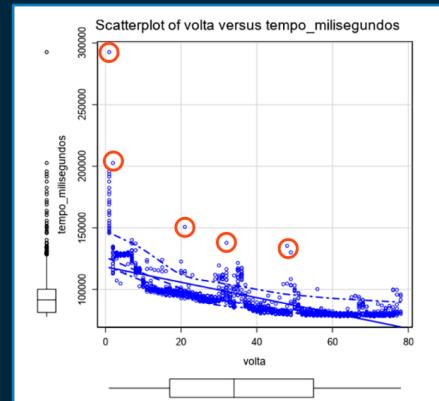
Spatial

34. Data Investigation question.

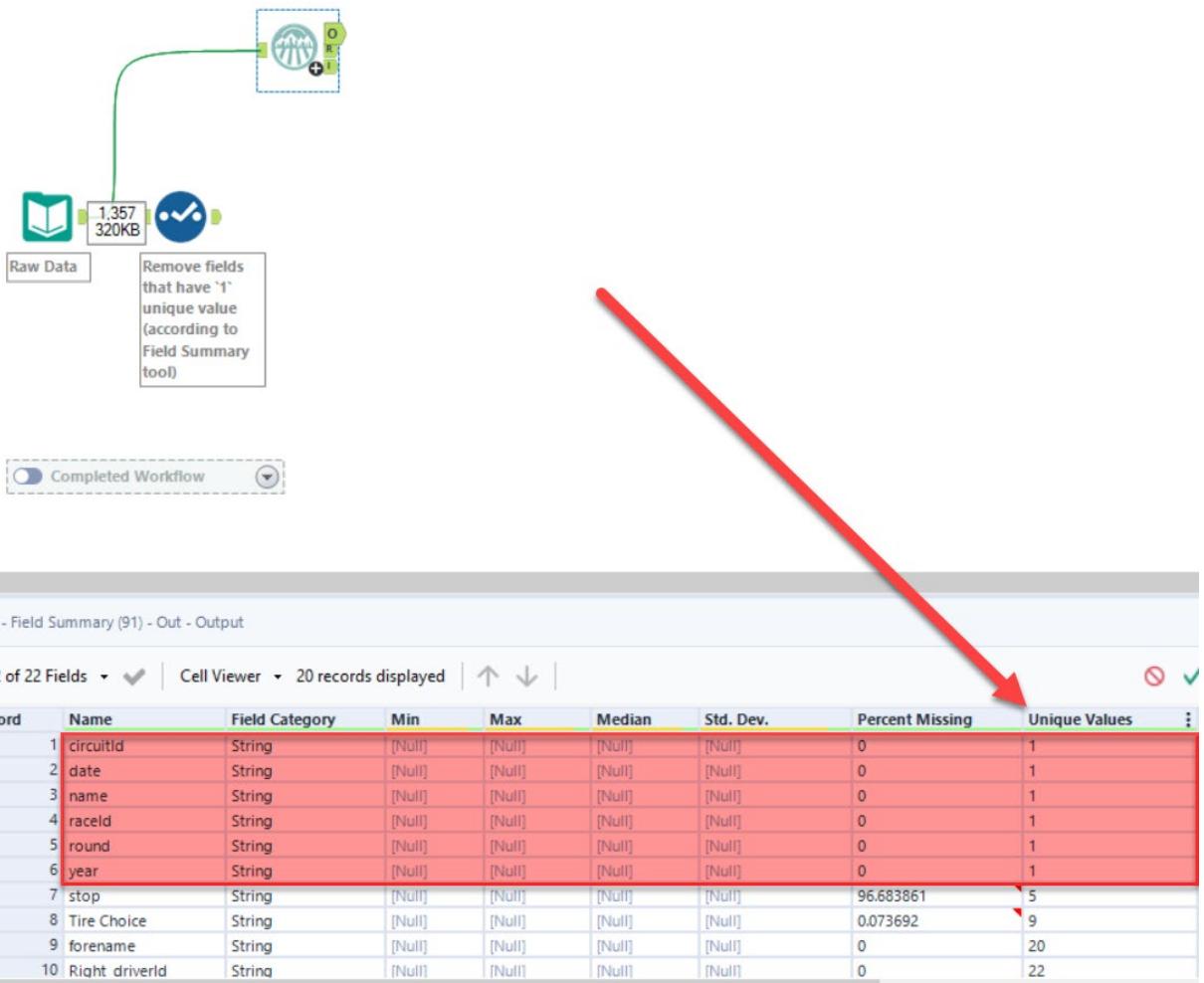
Question 3

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.



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.

Data Investigation Techniques

The screenshot shows a data investigation interface. At the top, there's a navigation bar with the title "Data Investigation Techniques". Below it is a toolbar with various icons. A main workspace contains a flow diagram with nodes: a book icon, a wavy line icon, and a circular icon with a gear and a magnifying glass. A dashed box highlights the circular icon. To the right is a table titled "Cell Viewer" showing 14 records displayed out of 22 fields. The table includes columns for Record, Name, Field Category, Min, Max, Median, Std. Dev., Percent Missing, Unique Values, and Mean. The rows show data for fields like Pit_time, lap, milliseconds, Pit Duration, Right_driverId, Time of Pit, Tire Choice, code, driverRef, forename, position, stop, surname, and time. The "Right_driverId" row is highlighted with a yellow background.

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	Pit_time	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.

When you find bias:

Oversample the minority

Undersample the majority

Use stratified sampling

38. Correlation

Correlation

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

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

Numeric Relationships Only



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

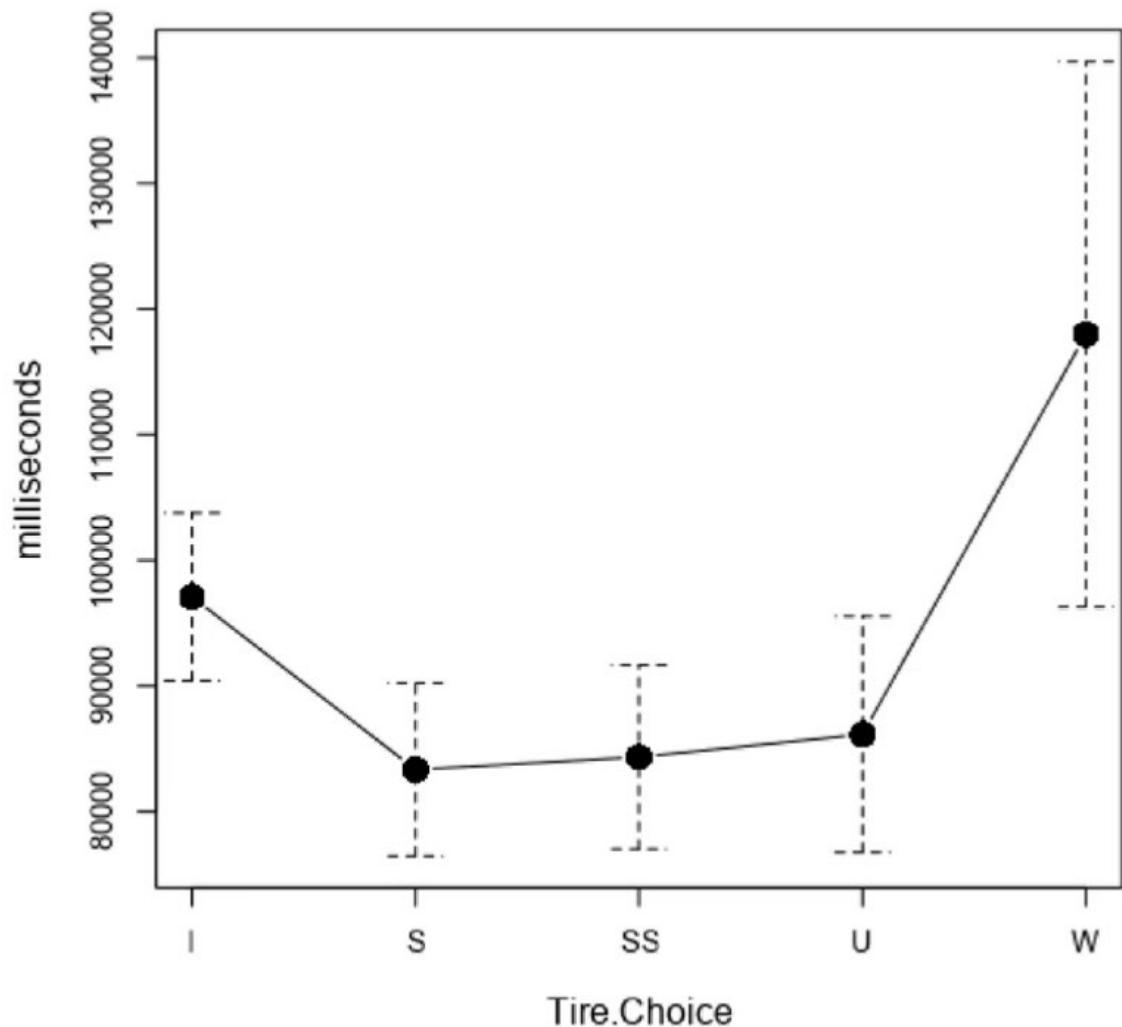
For relationships between continuous and categorical variables:



Example:

Graph

Plot of Means for milliseconds by Tire.Choice Levels



Example:

Configuration - Test of Means

Select the response field
milliseconds

Select the field with the group identifier
Tire Choice

The label for the control group (optional if there are only two groups)
I

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

The same information
in a different format.

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

For relationships between two categorical variables:



PAL	Frequency	0	0	0	0	7	7
	Percent	0	0	0	0	0.52	0.52
	Row Percent	0	0	0	0	100	-
	Column Percent	0	0	0	0	2.08	-
PER	Frequency	9	48	0	0	21	78
	Percent	0.66	3.54	0	0	1.55	5.75
	Row Percent	11.54	61.54	0	0	26.92	-
	Column Percent	3.13	33.1	0	0	6.25	-
RAI	Frequency	0	0	0	0	10	10
	Percent	0	0	0	0	0.74	0.74
	Row Percent	0	0	0	0	100	-
	Column Percent	0	0	0	0	2.98	-
RIC	Frequency	9	0	46	0	23	78
	Percent	0.66	0	3.39	0	1.69	5.75
	Row Percent	11.54	0	58.97	0	29.49	-
	Column Percent	3.13	0	18.33	0	6.85	-
ROS	Frequency	11	0	0	47	20	78
	Percent	0.81	0	0	3.46	1.47	5.75
	Row Percent	14.1	0	0	60.26	25.64	-
	Column Percent	3.82	0	0	13.95	5.95	-
SAI	Frequency	10	0				
	Percent	0.74	0				
	Row Percent	12.99	0				
	Column Percent	3.47	0				
VER	Frequency	19	3				
	Percent	1.4	0.22				
	Row Percent	55.88	8.82				
	Column Percent	6.6	2.07				
VET	Frequency	18	47				
	Percent	1.33	3.46				
	Row Percent	23.08	60.26				
	Column Percent	6.25	32.41				
WEH	Frequency	0	0				
	Percent	0	0				
	Row Percent	0	0				
	Column Percent	0	0				
Total	Frequency	288	145	251	337	336	1357
	Percent	2.22	10.69	18.5	24.83	24.76	100

A low Chi-squared value and a low p-value indicate a significant relationship between the categorical variables

Chi-squared = 1319.4283, df = 84, p-value < 0.0000

PREDICTOR VARIABLES:

- Strong relationship to target variable
- Unrelated to any other predictor variables
 - Being selective reduces collinearity

43. Feature selection

Feature Selection

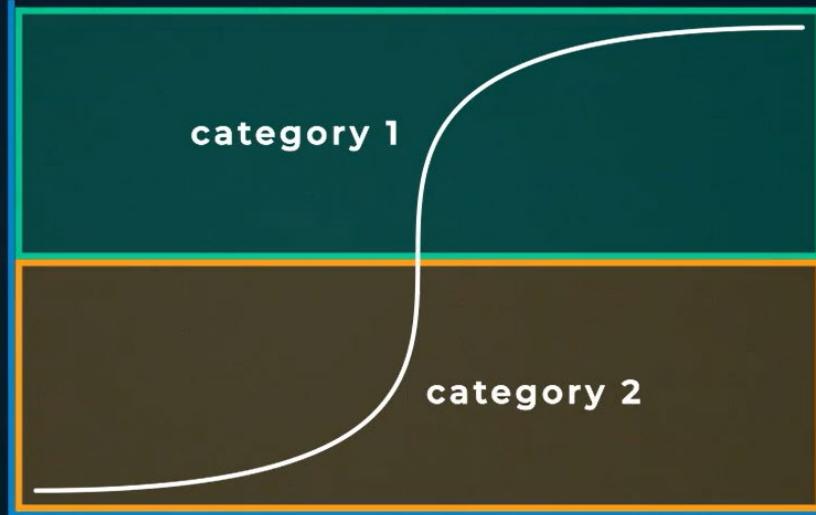
is choosing which variables to include as predictors in a modeling algorithm .

44. Logistic regression

Binary Classification



LOGISTIC
REGRESSION



Uses a regression to generate
ONE of TWO discrete class outputs

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



LOGISTIC
REGRESSION

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.

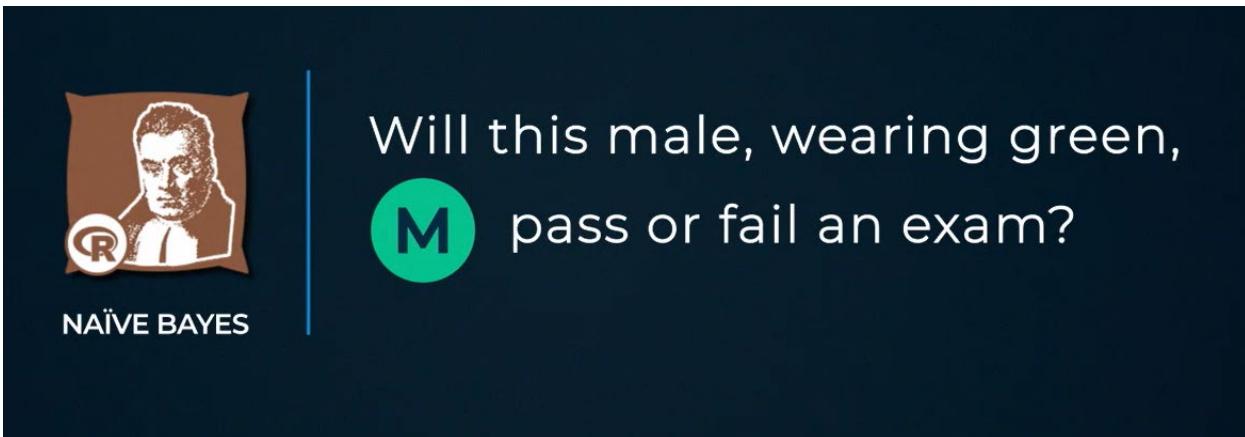
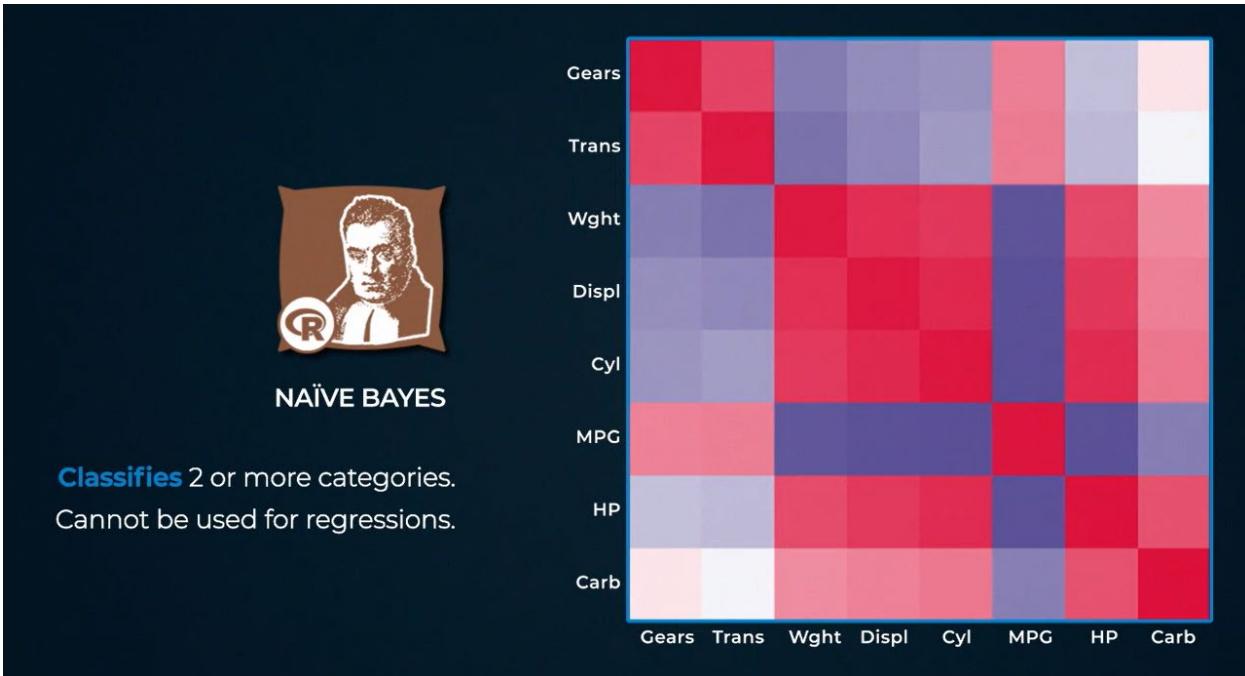


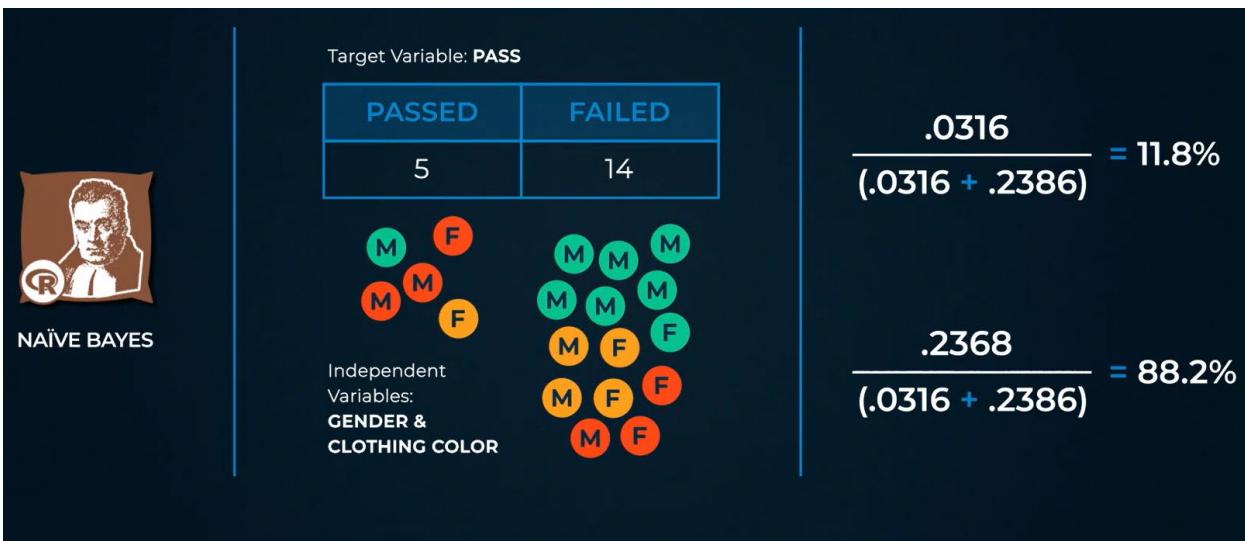
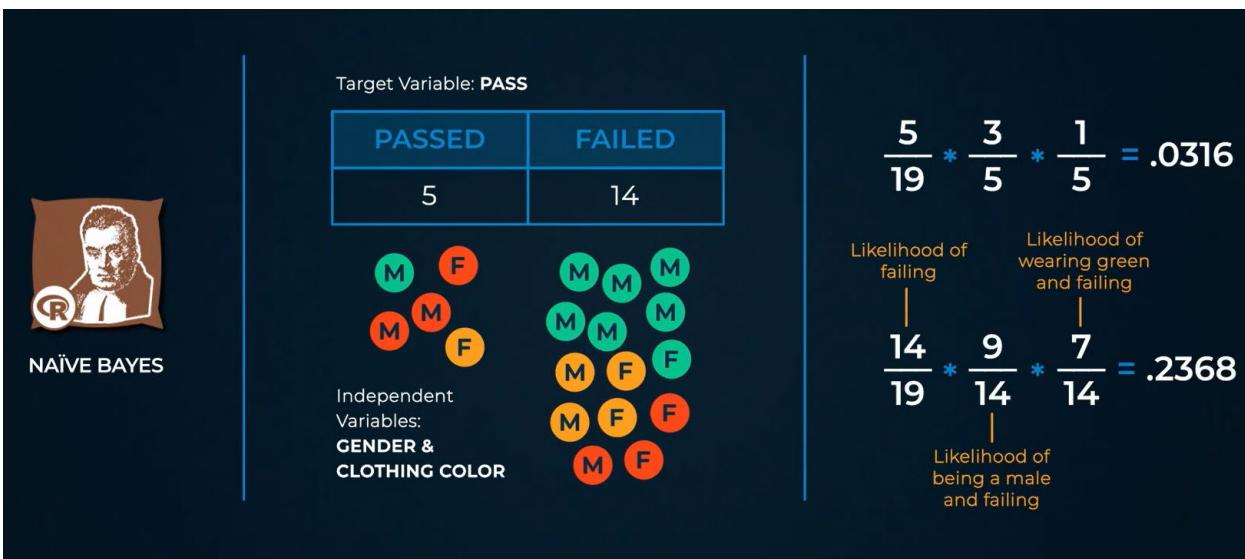
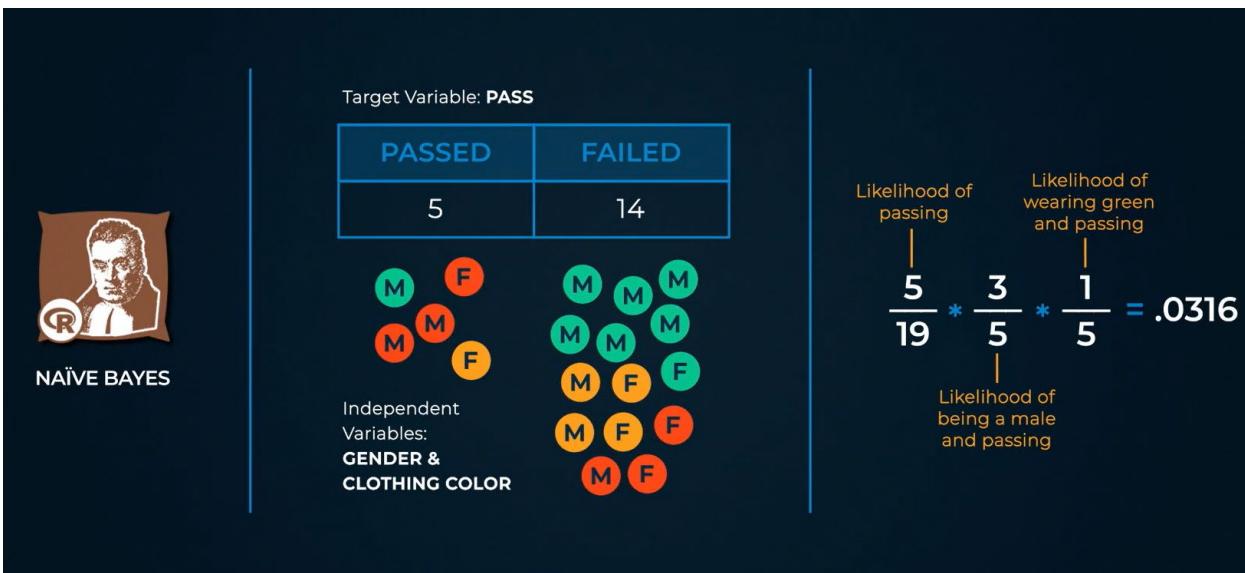
```

Record Report
1      Report for Linear Model Linear_Regression_188
2      Basic Summary
3      Call:
4      lm(formula = Wins ~ BatAge + H + SO, data = the.data)
5      Residuals:
6
7      Min       1Q   Median     3Q    Max
8      -22.15  -5.62    2.85   7.18  16.59
9
10     Coefficients:
11
12             Estimate Std. Error t value Pr(>|t|)
13 (Intercept) -218.25492  82.30169 -2.652  0.0137 *
14 BatAge        3.88566   1.68558  2.305  0.02973 *
15 H            0.08805   0.03084  2.855  0.00853 **
16 SO           0.05329   0.02420  2.202  0.03708 *
17
18     Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
19
20     Residual standard error: 9.6515 on 25 degrees of freedom
21     Multiple R-squared: 0.3649, Adjusted R-Squared: 0.2887
22     F-statistic: 4.788 on 3 and 25 degrees of freedom (DF), p-value 0.009036
23
24     Type II ANOVA Analysis
25
26     Response: Wins
27
28             Sum Sq DF   F value Pr(>F)
29 BatAge        495.02  1     5.31  0.02973 *
30 H            759.22  1     8.15  0.00853 **
31 SO           451.84  1     4.85  0.03708 *
32 Residuals   2328.81 25
33
34     Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

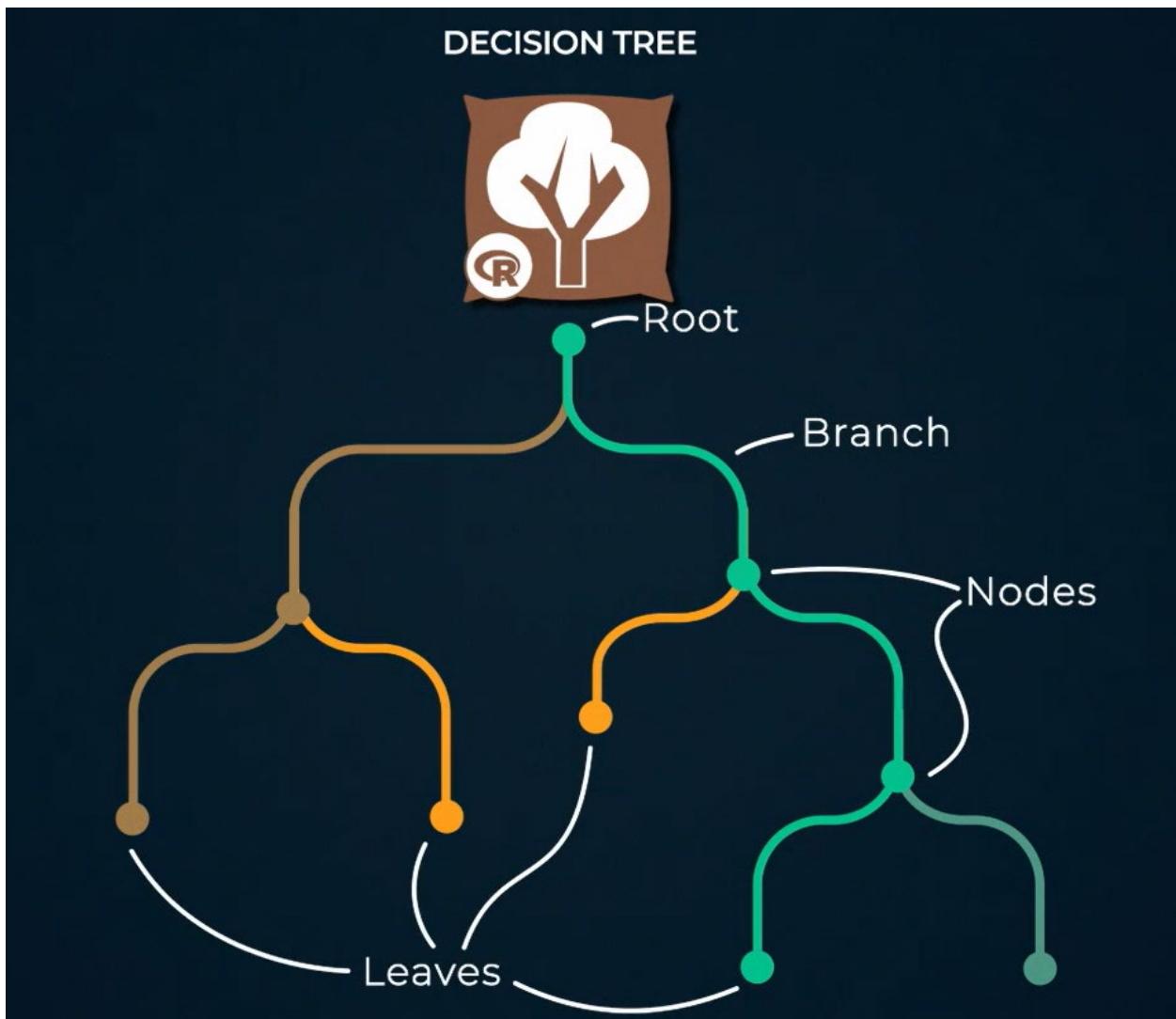
```

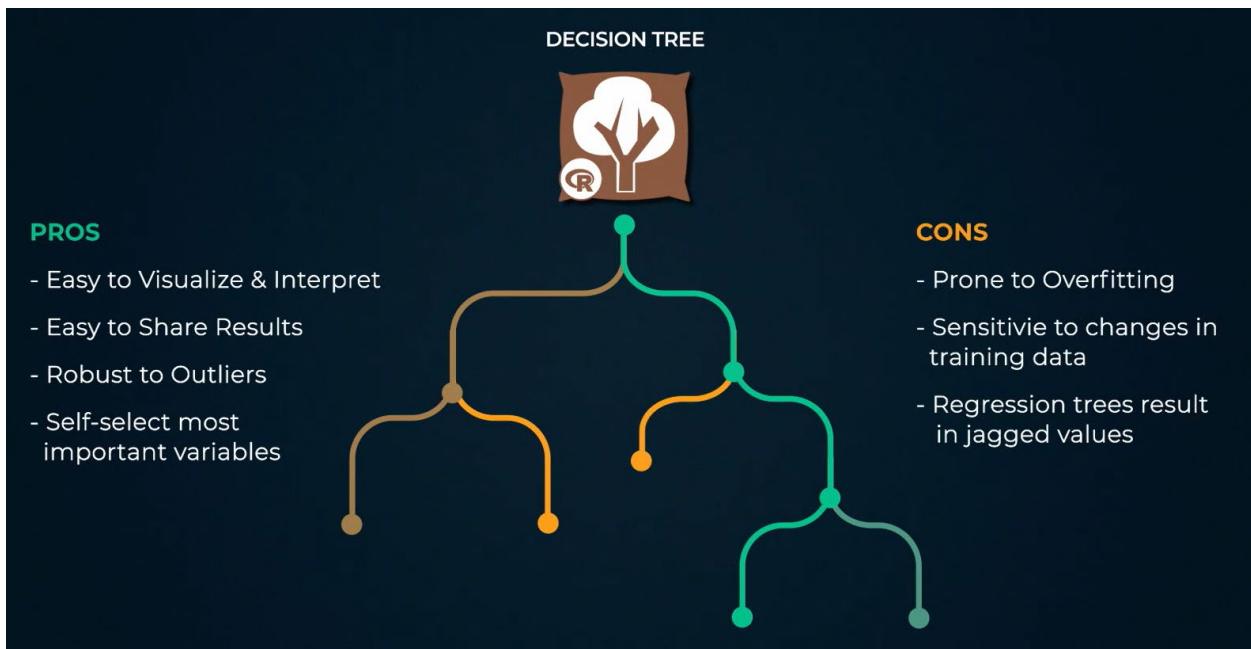
46. Naïve Bayes. Called naïve because it assumes all predictor (i.e., independent) variables are totally independent.





47. Decision Tree





DECISION TREE



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)

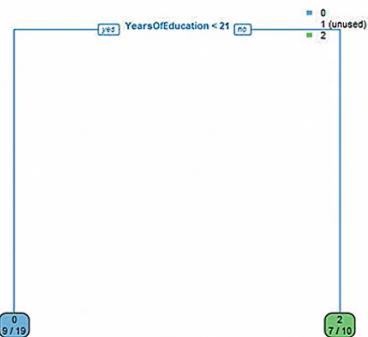
2) YearsOfEducation< 20.5 19 10 0 (0.47368421 0.47368421 0.05263158) *

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

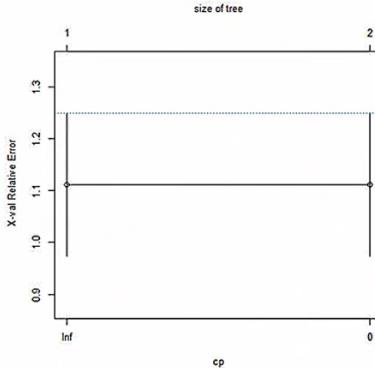
Each split is “associated” with a label.

Were you to apply those labels after the first grouping, your model would have this accuracy.

Tree Plot



Pruning Plot



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

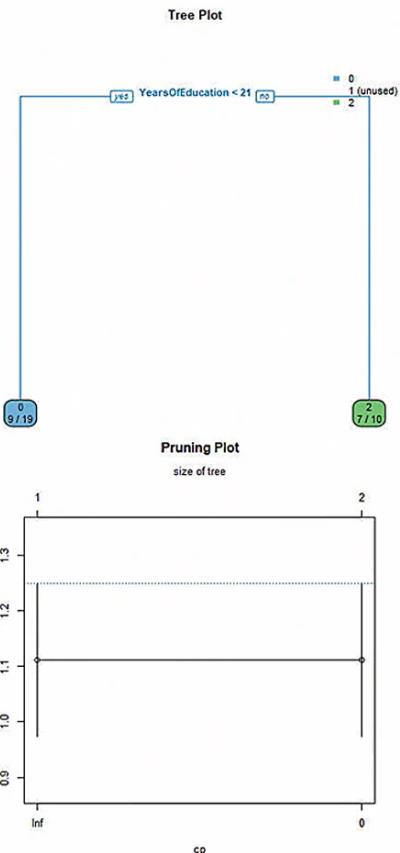
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.27778	0	1.00000	1.1111	0.13841
2	0.00000	1	0.72222	1.1111	0.13841

Leaf Summary

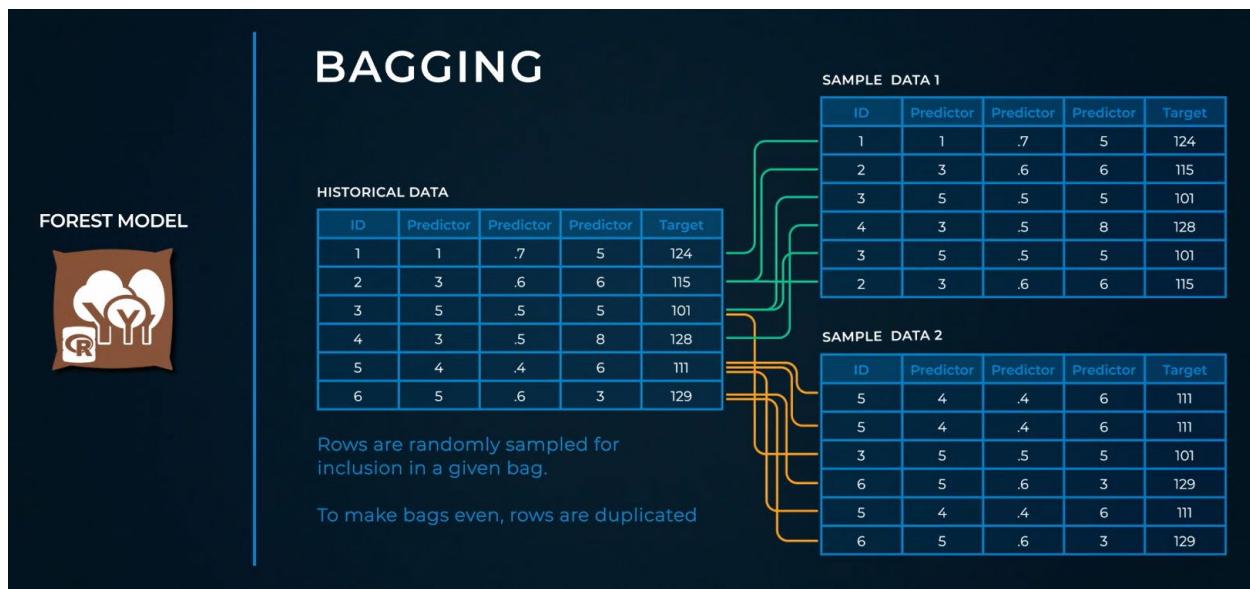
node), split, n, loss, yval, (yprob)
* denotes terminal node
1) root 29 18 0 (0.37931034 0.34482759 0.27586207)
2) YearsOfEducation< 20.5 19 10 0 (0.47368421 0.47368421 0.05263158) *
3) YearsOfEducation>=20.5 10 3 2 (0.20000000 0.10000000 0.70000000) *

Add these values and compare to the xError.

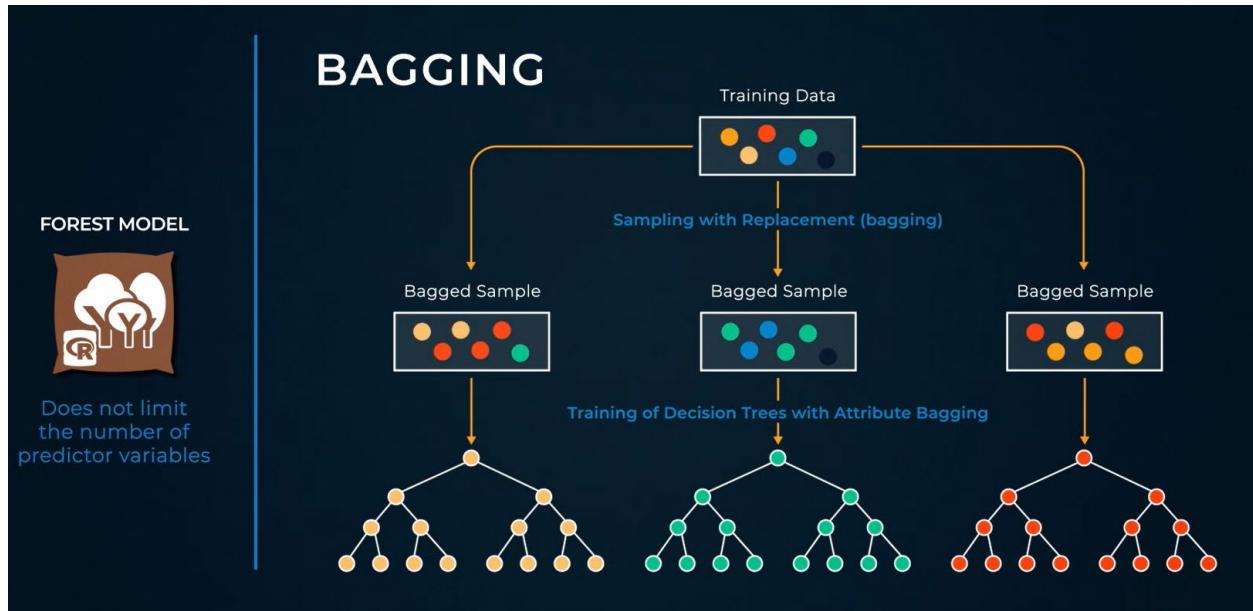
If sum > xError, then prune that branch.



48. Forest model



There will be some duplicated
and some excluded records



49. Boosted model

**FOREST
MODEL**



Creates trees using
RANDOM BAGGING

**BOOSTED
MODEL**



Creates trees
IN SEQUENCE

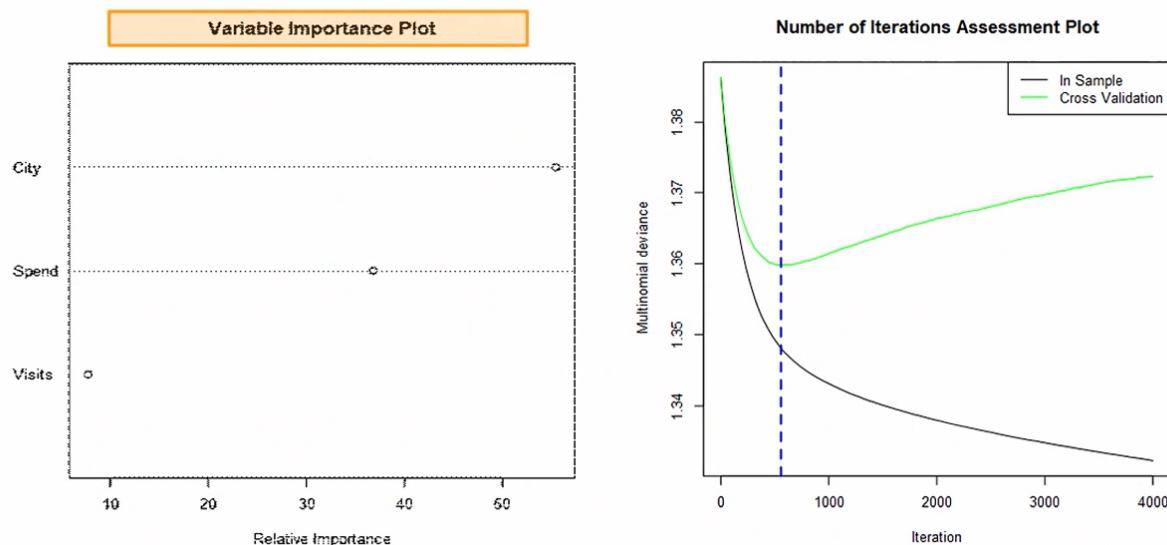
Report for Boosted Model Boosted

Basic Summary:

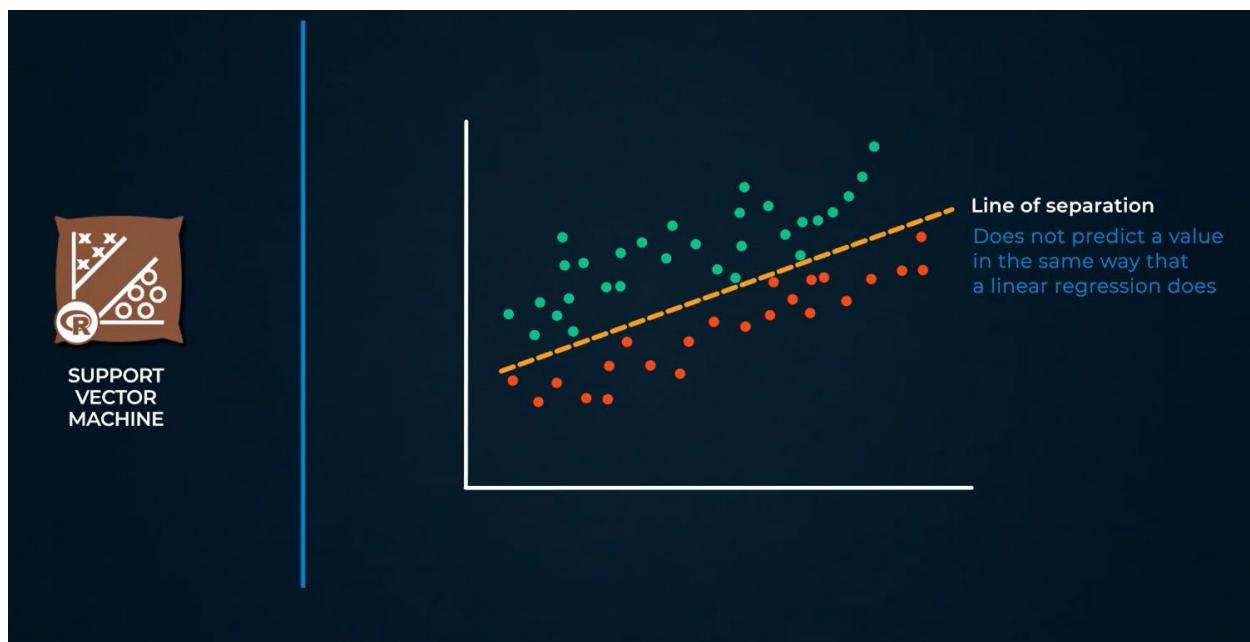
Loss function distribution: Multinomial

Total number of trees used: 4000

Best number of trees based on 5-fold cross validation: 557

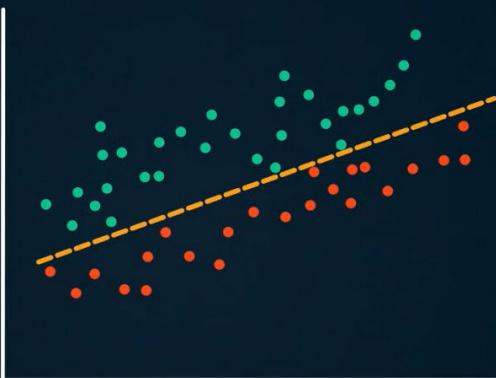


50. Support Vector Machine





SUPPORT VECTOR MACHINE



PROS

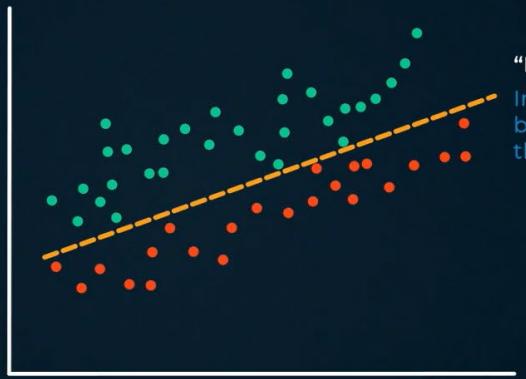
- Resistant to overfitting
- Can outperform other algorithms on difficult classifications

CONS

- Computationally expensive
- Require fine-tuning

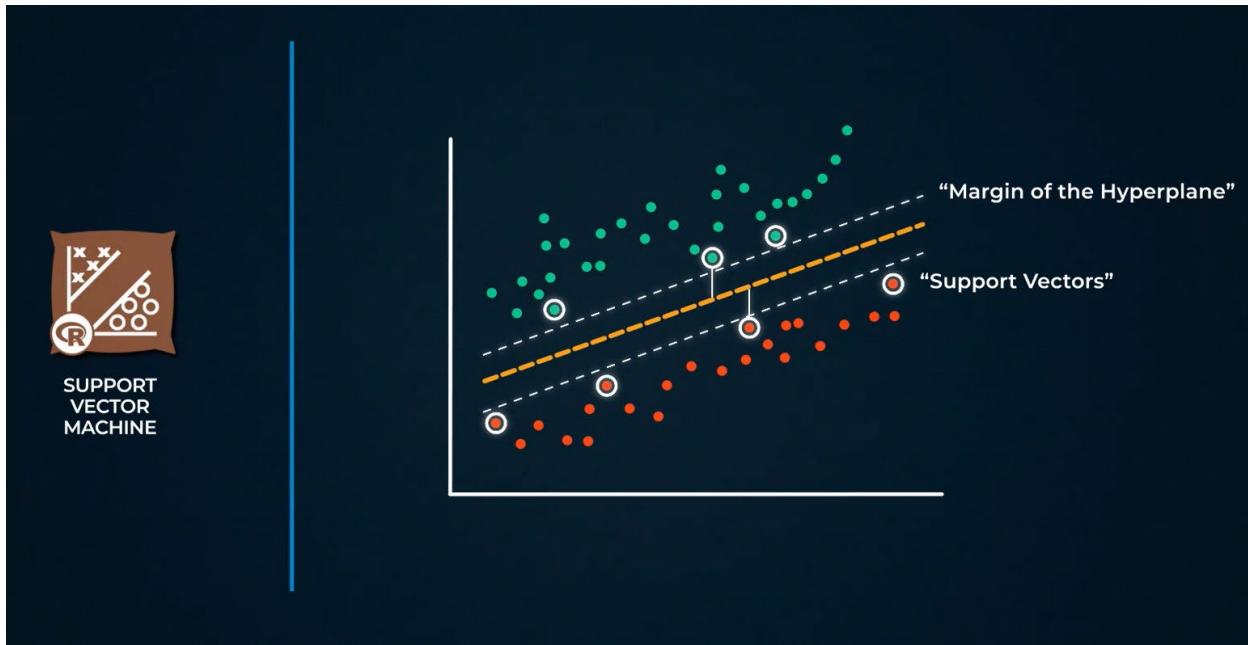


SUPPORT VECTOR MACHINE



"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 + 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, 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

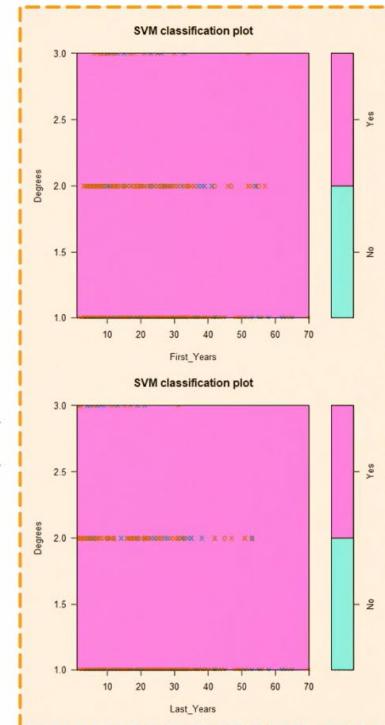
Confusion Matrix

	No	Yes
No	430	100
Yes	173	400

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 model. 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

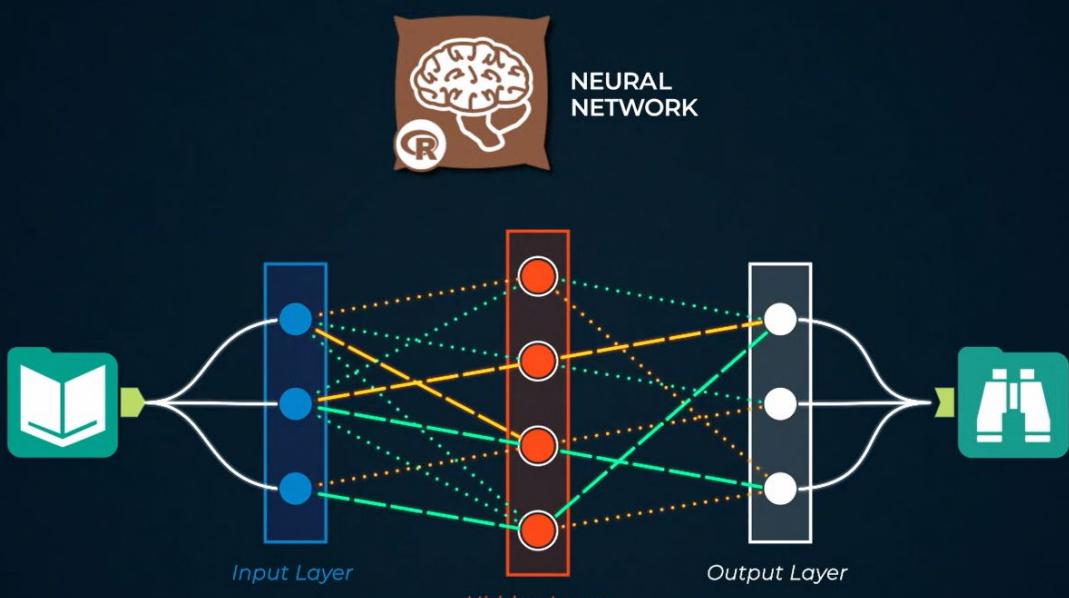
PROS

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

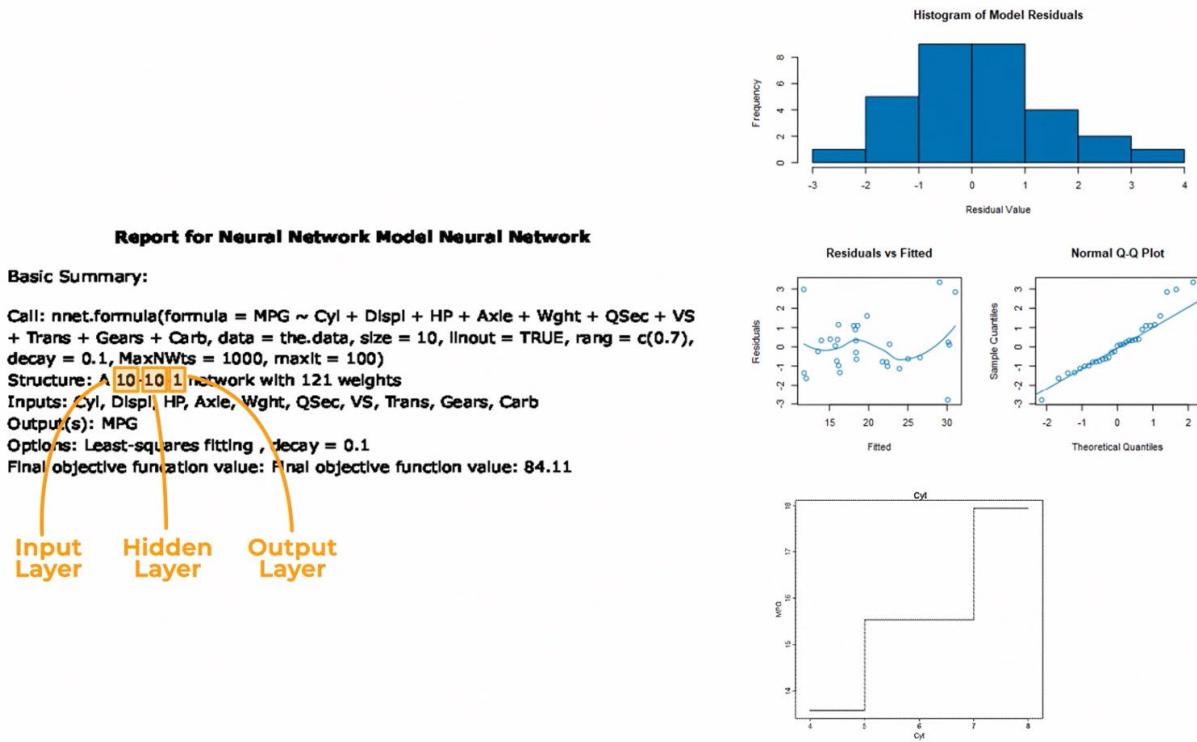


CONS

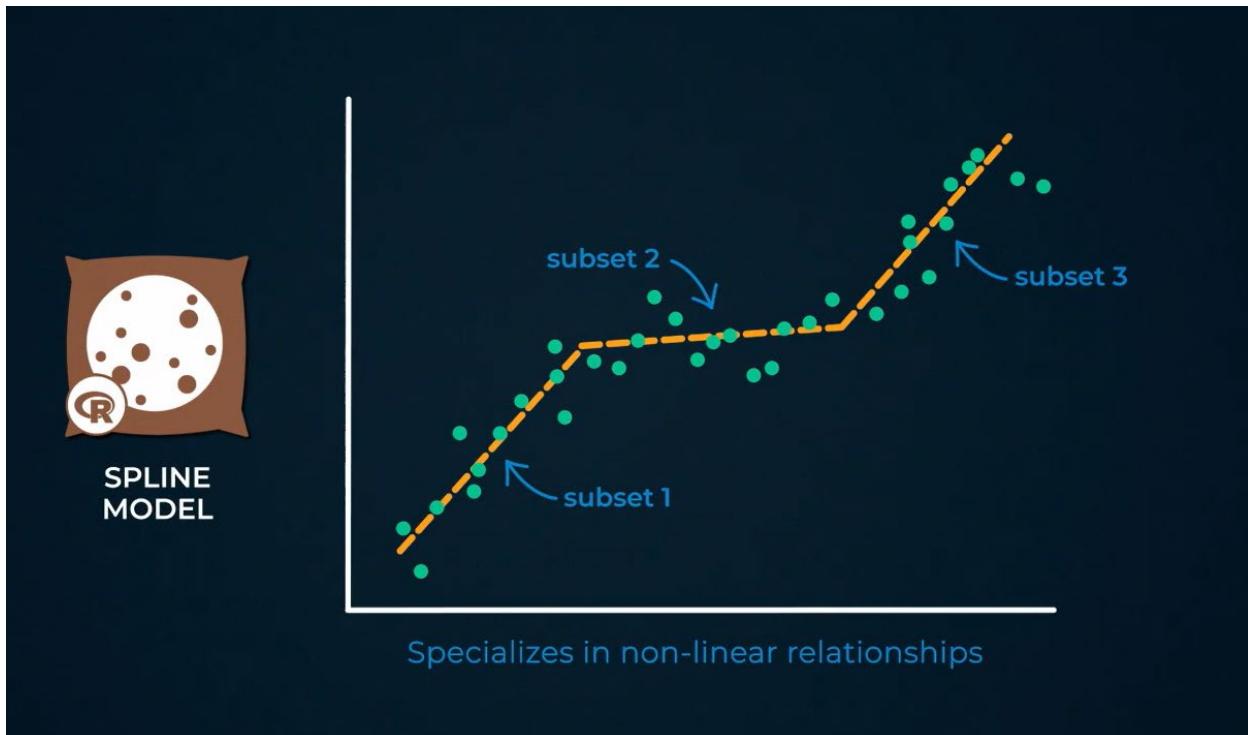
- Difficult to explain
- Sensitive to missing values
- Sensitive to outliers
- Time consuming to create



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

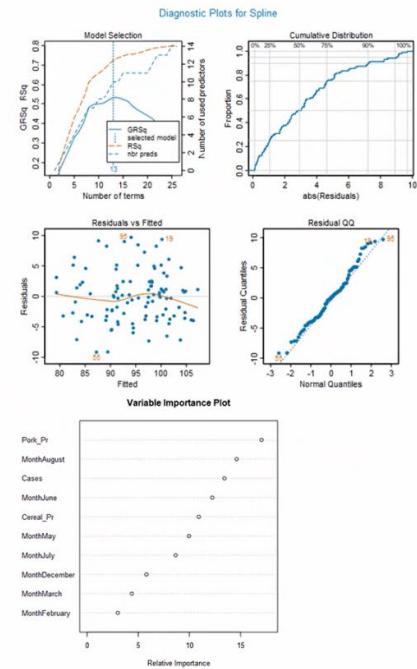
Selected 13 of 26 terms, and 10 of 20 predictors

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

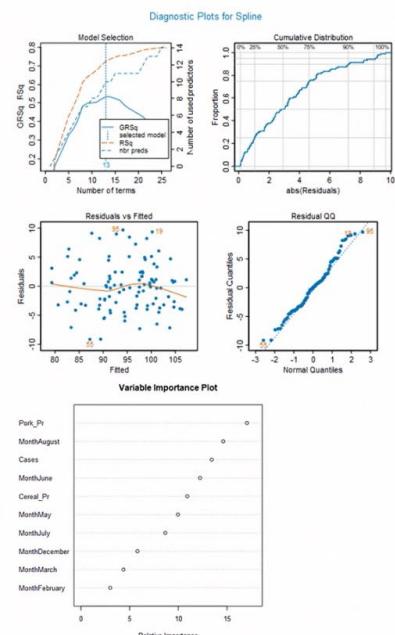
Selected 13 of 26 terms, and 10 of 20 predictors

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

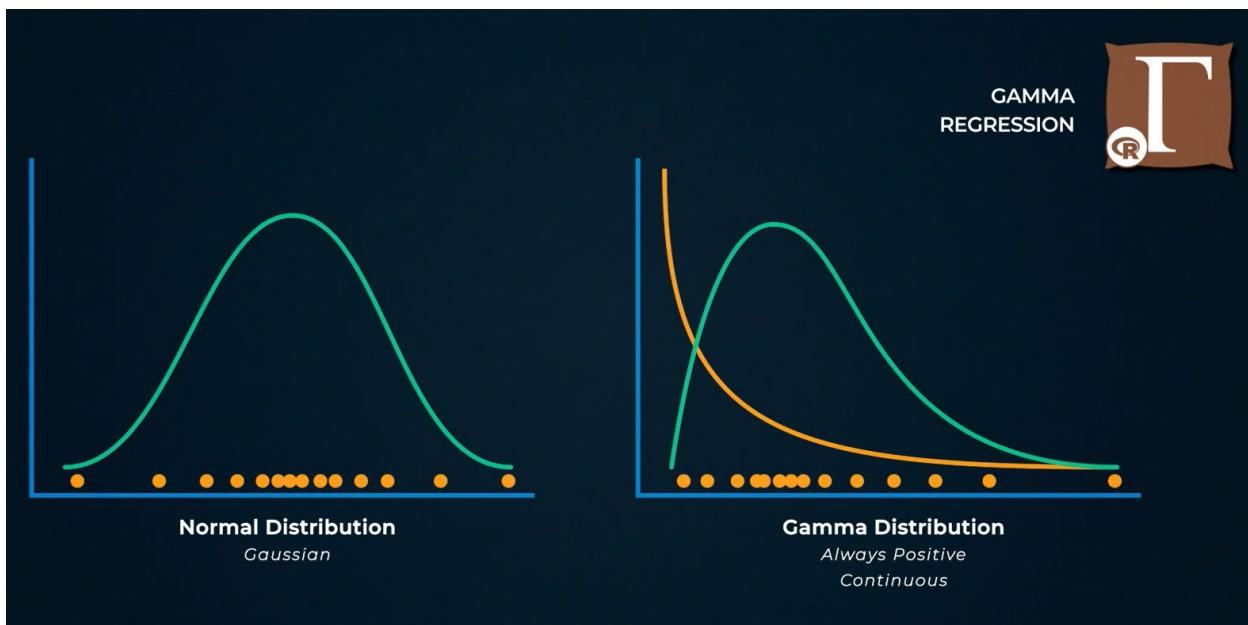
GCV 30.29 RSS 1818 GRSq 0.5323 RSq 0.7249

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

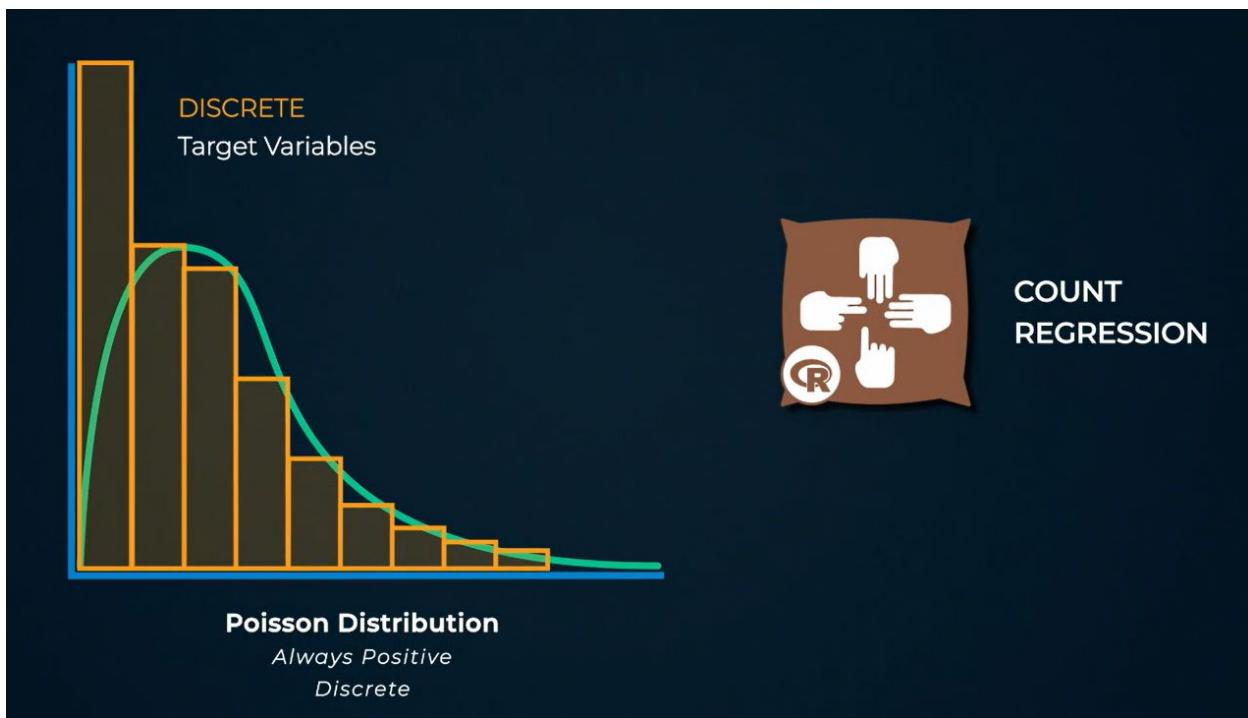
GLM Model McFadden R-Squared: 0.7249



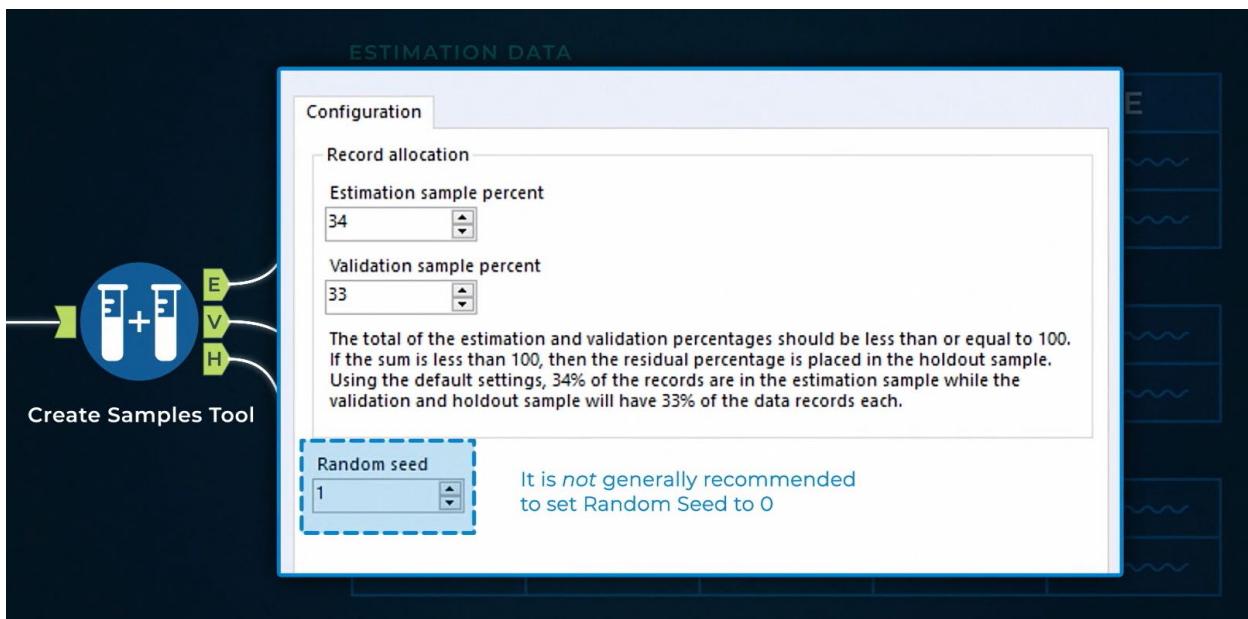
53. Gamma regression



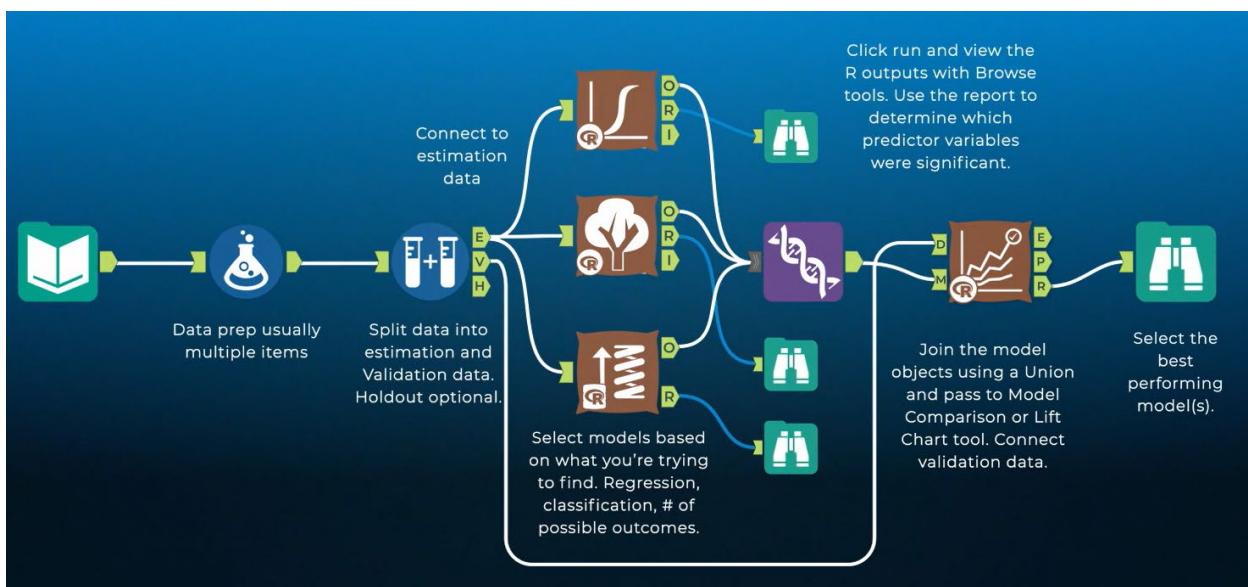
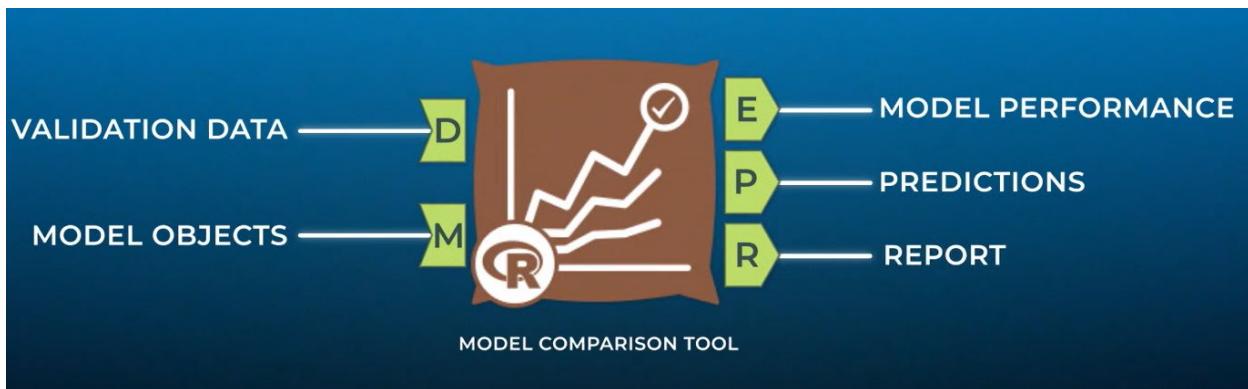
54. Count Regression

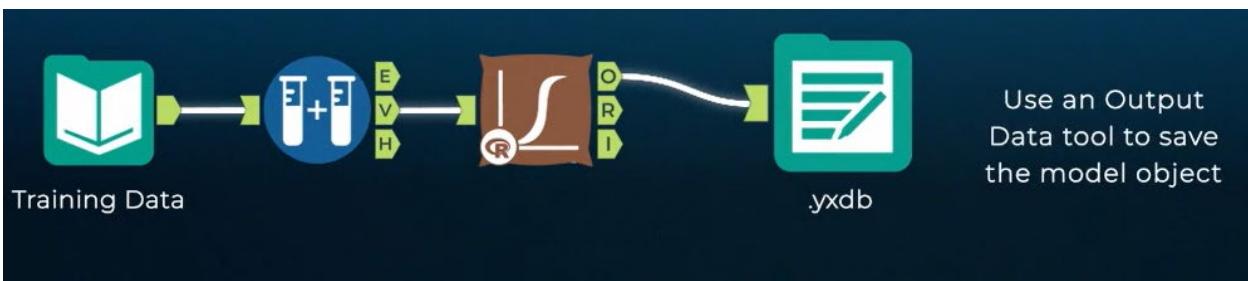
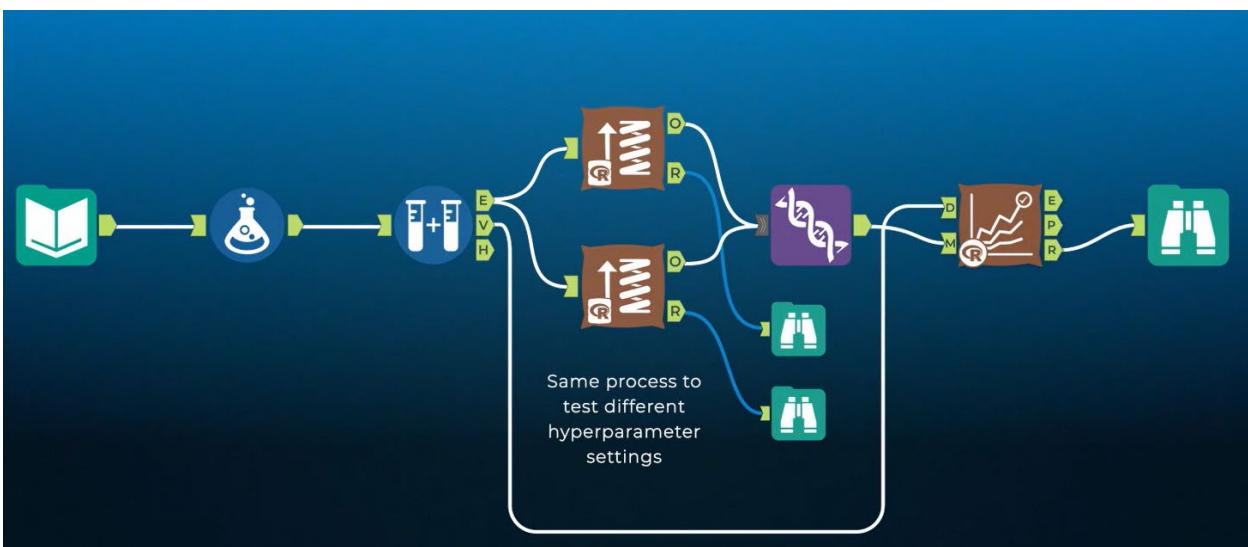
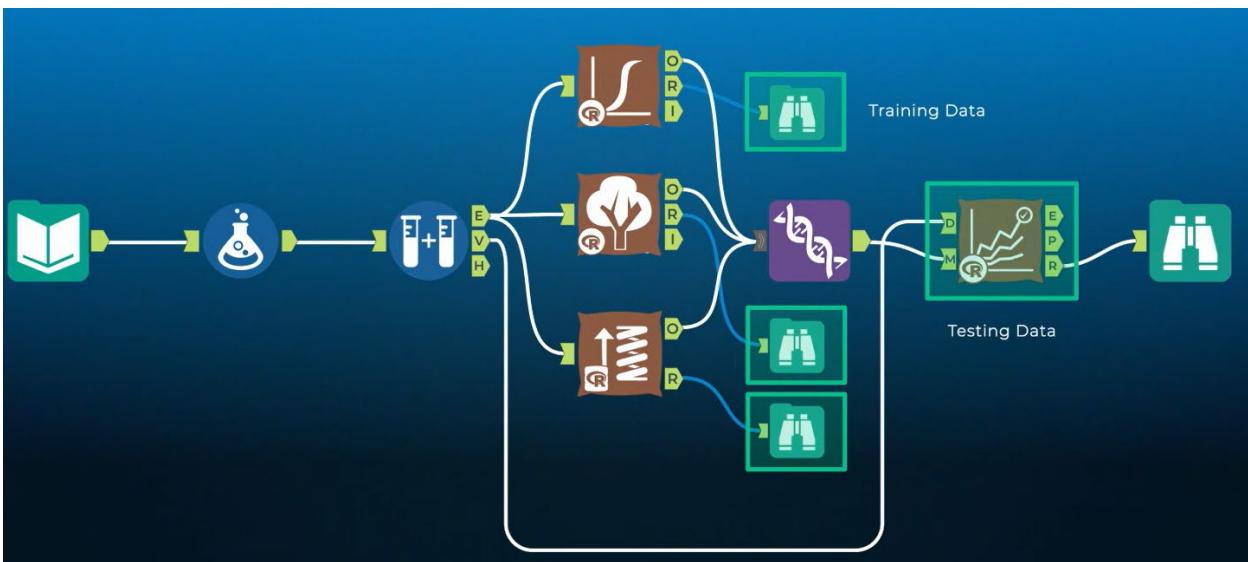


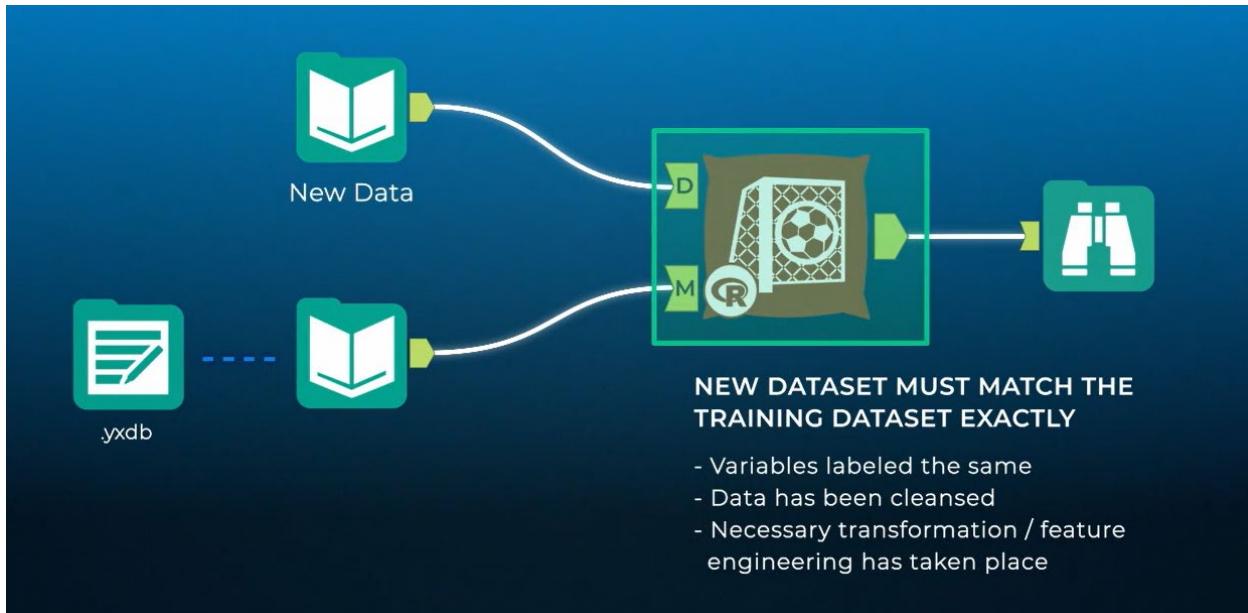
55. Create Samples. If you use a Random seed of '0', then you make it impossible to generate the same results since randomly selected records will be output on each run of the workflow.



56. Model Comparison and Score







57. Predictive Questions

Question 1

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

Support Vector Machine

Logistic Regression

Neural Network

Gamma Regression

Boosted Model

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.

- 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.

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

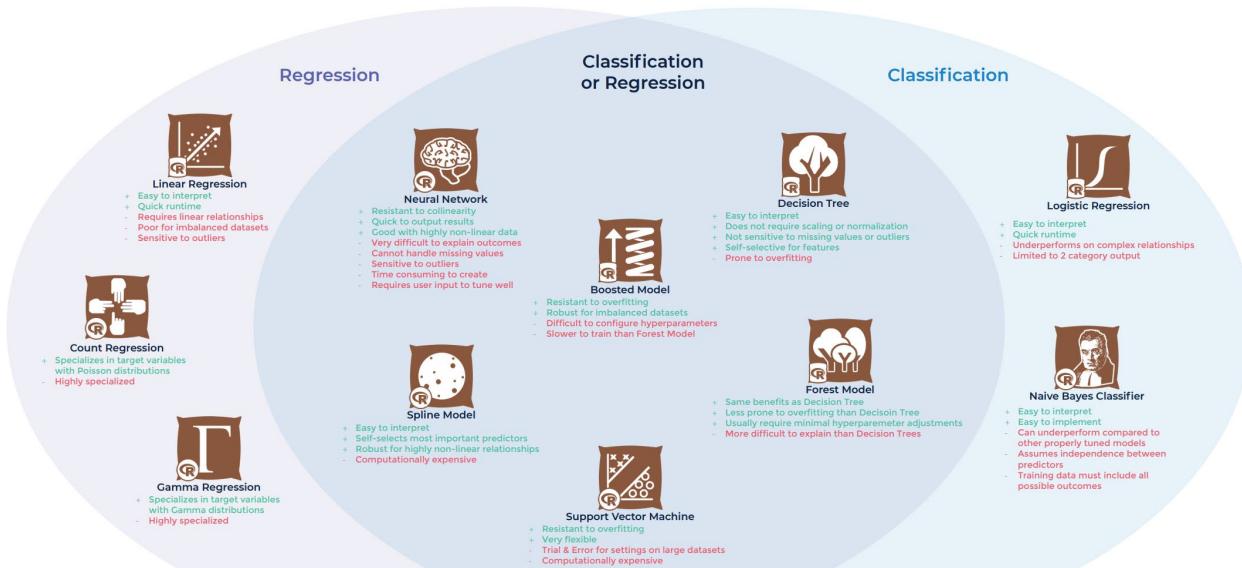
Question 10

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

SELECTING A PREDICTIVE MODELING ALGORITHM



59. Principal Component Analysis (PCA)

PRINCIPAL COMPONENT ANALYSIS



VAR 1	VAR 2	VAR 3	VAR 4	VAR 5	VAR 6

PC 1	PC 2	PC 3

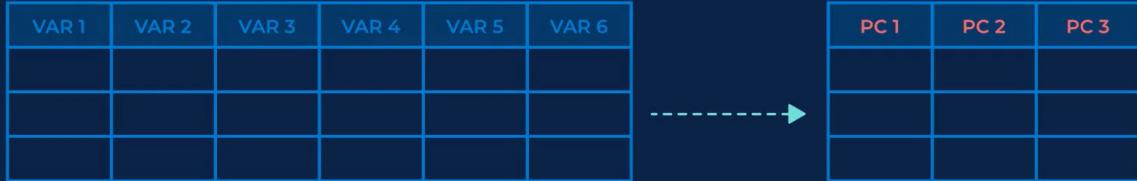
BENEFITS:

- 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



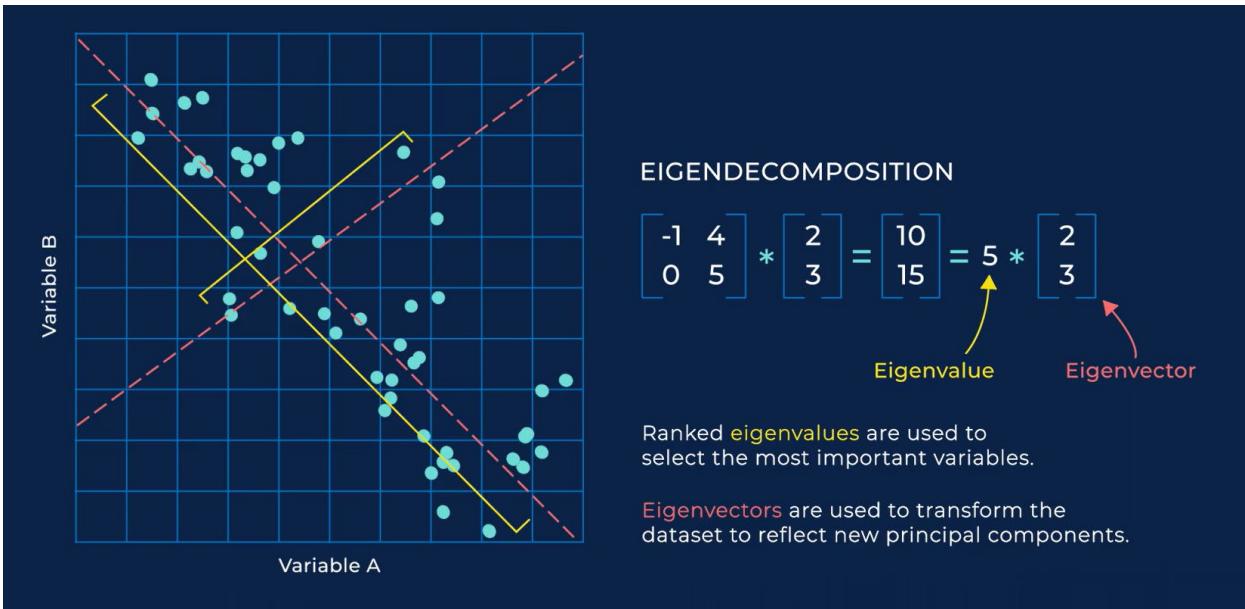
New variables are created from weighted
combinations of existing variables.

	A	B
A	0.67	0.55
B	0.55	0.25

COVARIANCE MATRIX

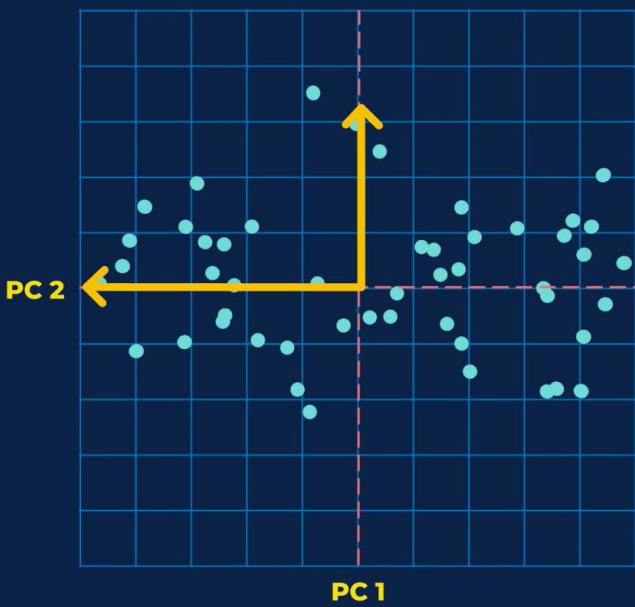
Describes how related
two variables are

Not the same as a
correlation matrix



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

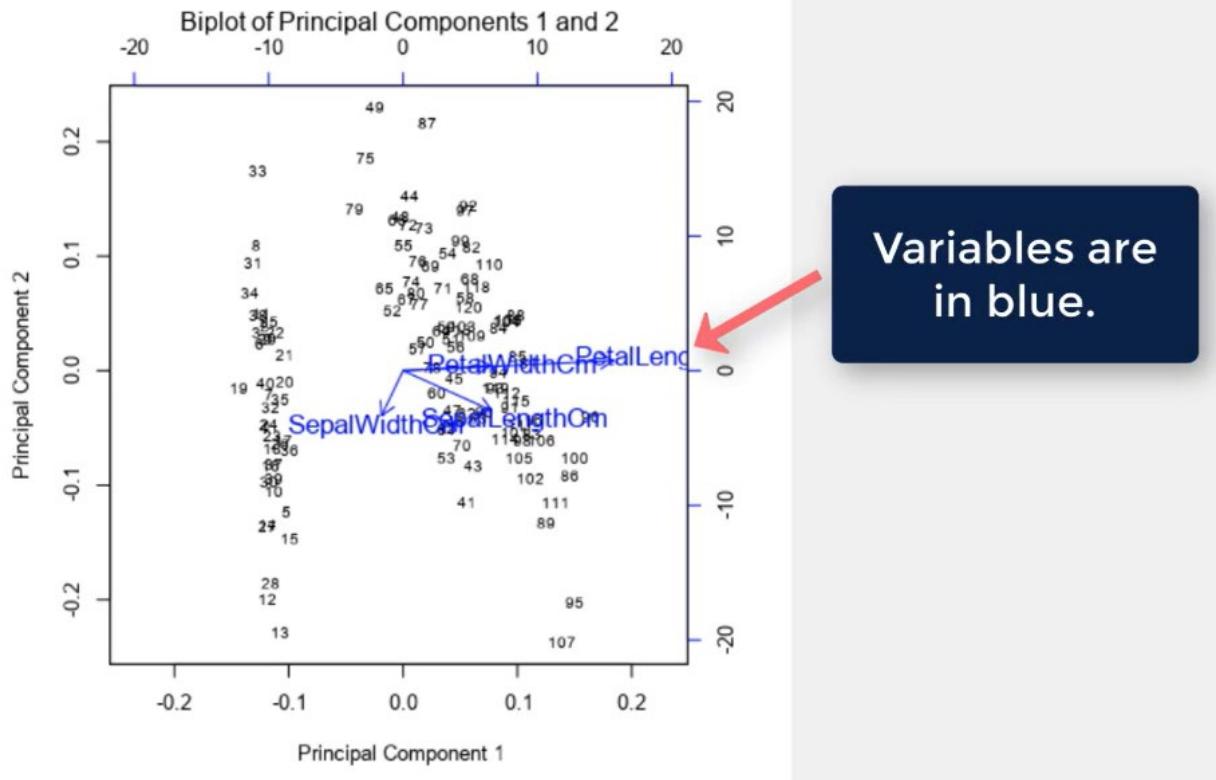
PROJECTION TO NEW COMPONENTS



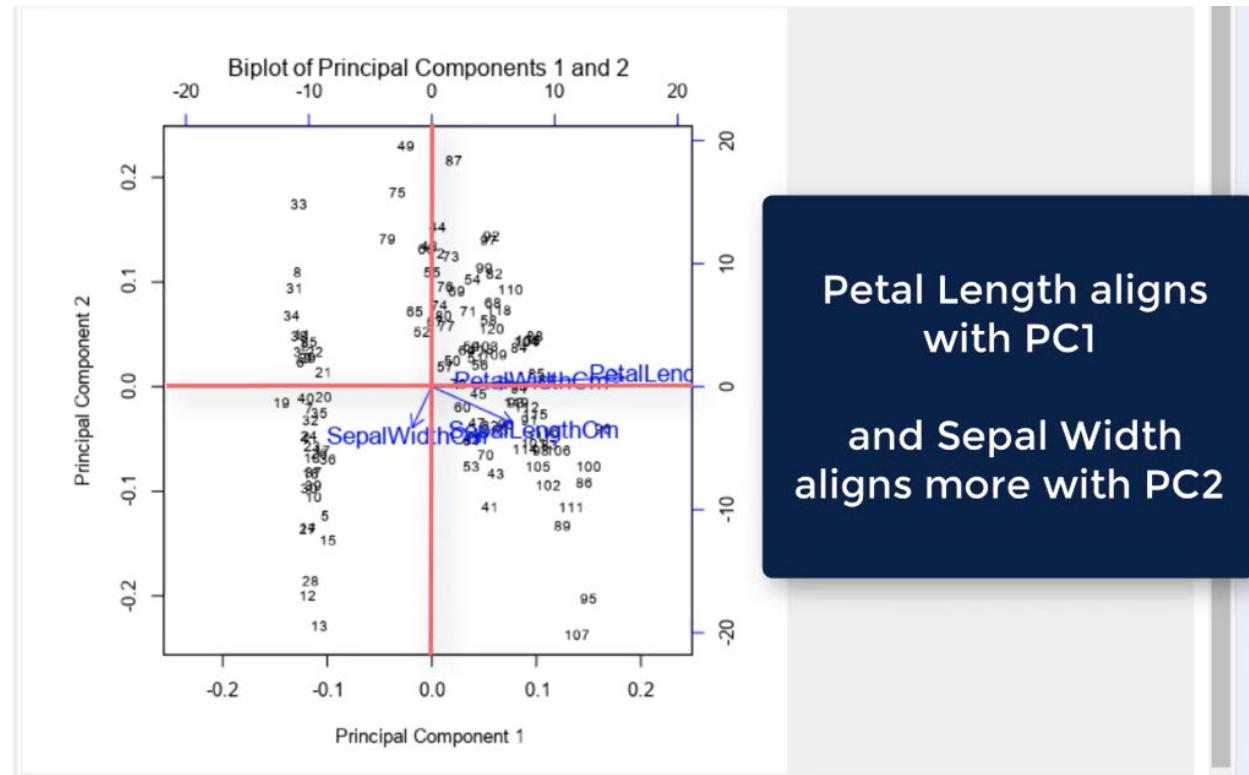
TRANSFORMATION

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.

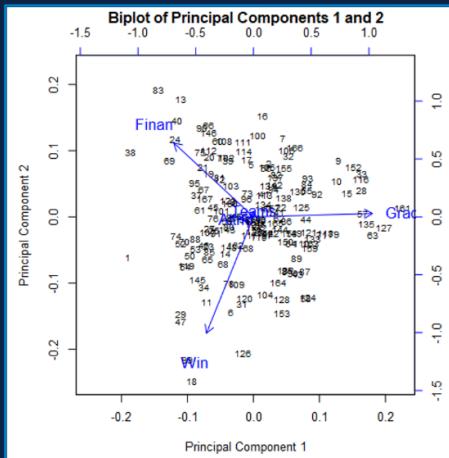
QUESTION 1

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
- Interpretability is lost
- Noisy data is interpreted as important



Text Input (2) - Configuration

Find Nearest.yxmd X Neighbors Workflow.yxmd* X + ...

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
4	35	4.9	3.1	1.5	0.1
5	38	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 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:



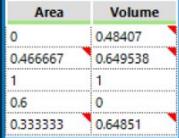
Results - Join (6) - Out - Join

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?

● 

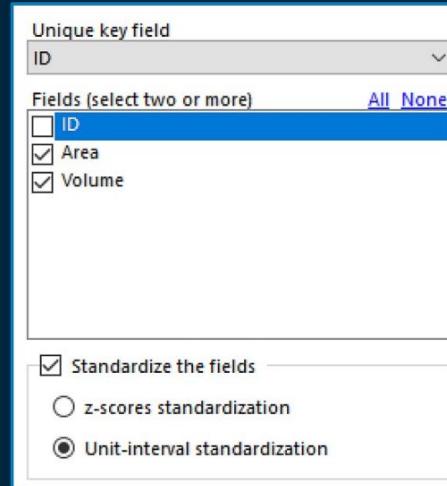
Area	Volume
0	0.48407
0.466667	0.649538
1	1
0.6	0
0.333333	0.64851

● 

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



Unique key field
ID

Fields (select two or more) [All](#) [None](#)

ID Area Volume

Standardize the fields

z-scores standardization Unit-interval standardization

Question 2

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

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

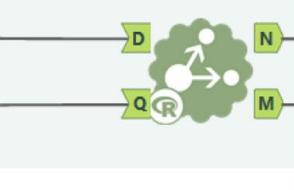
Question 3

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

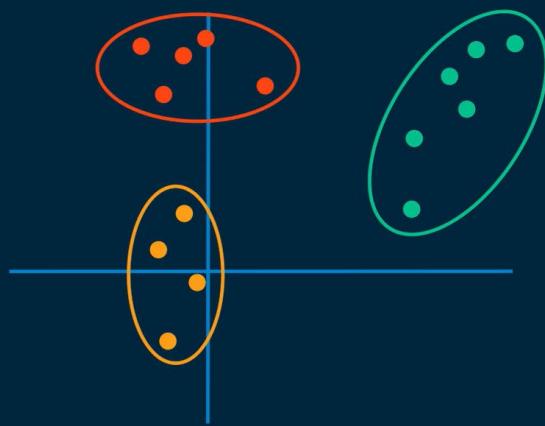
ID	Area	Volume
5	15.2	1730
6	21.8	1695



61. Clusters

CLUSTERS

Groupings of similar datapoints and separation of dissimilar datapoints.



STEPS

User inputs K = 2

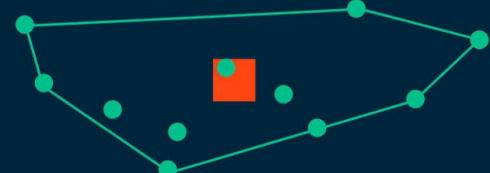
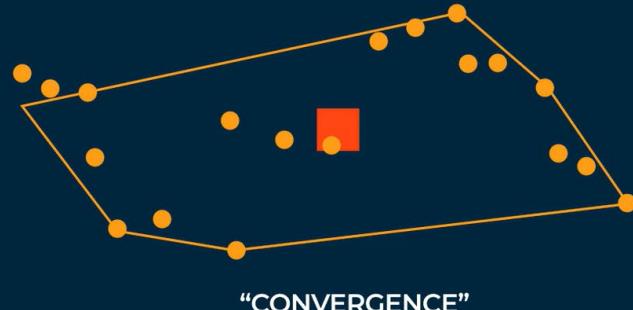
2 randomly selected points are assigned as centroids.

The distance of all other points are calculated 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



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

K-Centroids Cluster Analysis (2) - Configuration

Configuration Plot Options Graphics Options

Solution name: Iris

Fields (Select two or more):
Id (checkbox checked), SepalLengthCm (checkbox checked), SepalWidthCm (checkbox checked), PetalLengthCm (checkbox checked), PetalWidthCm (checkbox checked)

All None

Standardize the fields... (checkbox)

Clustering method:
K-Means (radio button checked), K-Medians (radio button), Neural Gas (radio button)

Number of clusters: 2

Number of starting seeds: 10

Select Variables
Only numeric columns are displayed
Minimum of 2 variables selected
Cannot contain Null values
Sensitive to outliers

Z-Score

The screenshot shows a software interface for creating a new column. The 'Output Column' section has a dropdown menu open, with 'Stdzsd' selected. Below it, the formula $([Value] - [Mean]) / [SD]$ is displayed. To the left of the formula are icons for a function editor (fx), a mean symbol (X), a scatter plot (Scatter), and a histogram (Histogram). At the bottom, there are dropdown menus for 'Data type' set to 'Double' and 'Size' set to '8'.

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

Unit Interval

The screenshot shows a software interface for defining a new column. The 'Output Column' field contains the name 'Stndzd'. Below it, the formula $([Value] - [Min_Val])/([Max_Val] - [Min_Val])$ is displayed. To the left of the formula are icons for a function, a variable, a file, and a clipboard. At the bottom, the 'Data type:' dropdown is set to 'Double' and the 'Size:' input field is set to '8'.

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.

K-Centroids Diagnostics (4) - Configuration

Configuration Graphics Options

Fields (select two or more)

Id
 SepalLengthCm
 SepalWidthCm
 PetalLengthCm
 PetalWidthCm

Standardize the fields...

Clustering method

K-Means
 K-Medians
 Neural Gas

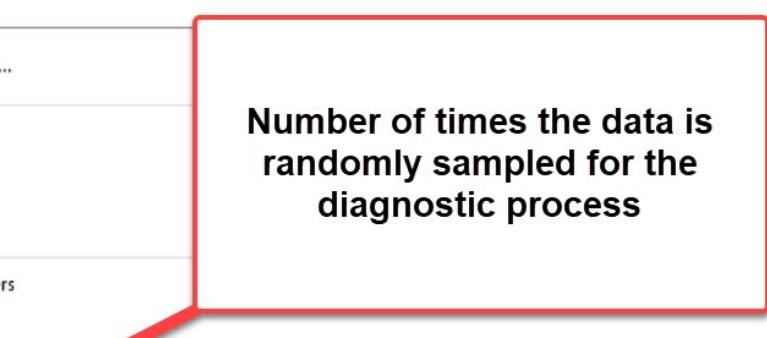
Minimum number of clusters
2

Maximum number of clusters
4

Bootstrap replicates
50

Number of starting seeds
3

Number of times the data is randomly sampled for the diagnostic process



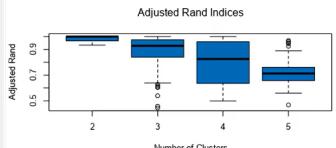
Bootstrap Replication

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.

K-Means Cluster Assessment Report					
Summary Statistics	2	3	4	5	
Adjusted Rand Indices:					
Minimum	0.933446	0.4442	0.498994	0.467505	
1st Quartile	0.966441	0.85504	0.635886	0.657455	
Median	1	0.927988	0.823531	0.712822	
Mean	0.98192	0.872852	0.787437	0.720324	
3rd Quartile	1	0.975191	0.959358	0.756627	
Maximum	1	1	1	1	0.969299
Calinski-Harabasz Indices:					
Minimum	372.2527	209.0601	277.7686	266.2095	
1st Quartile	388.2438	424.65	375.3327	340.9549	
Median	393.3448	434.2265	400.9253	355.1047	
Mean	391.6379	423.7289	384.9446	353.1681	
3rd Quartile	395.7433	438.1593	407.4645	372.1192	
Maximum	398.8446	445.1008	416.8787	388.2493	
Plots					
Adjusted Rand Indices					
					

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

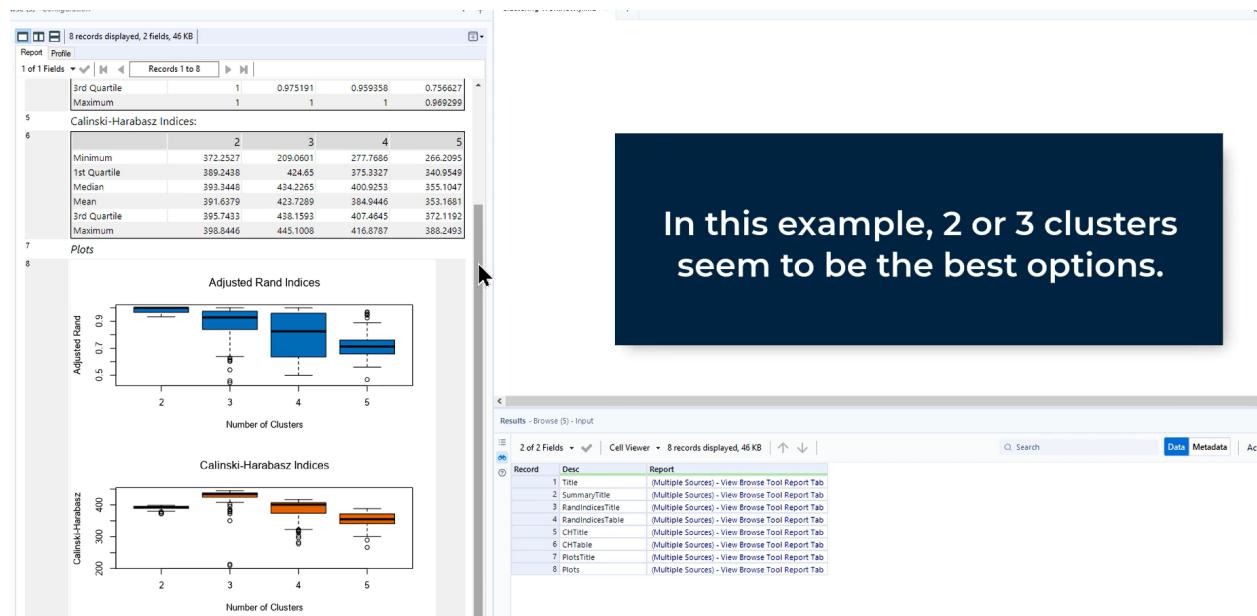
Definition of the clusters

-similarity of like datapoints vs
differences between clusters

Ratio:

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

Higher values are desirable



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

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.

- K-Means uses mean values to calculate centroids
- K-Medians uses median values to calculate centroids
- 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.

- Compressed box plots with high means on the Calinski-Harabasz Index
- Larger box plots on either index
- 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

CLUSTERING

ID	AREA	VISITORS	GROUP
52	36.5	556.2	●
45	19.9	89.6	●
19	120.2	1256.9	●
18	102.1	1006.3	●
94	71	742.7	●
37	97.5	658.0	●
71	49.6	733.8	●
16	99.3	155.8	●

new data point

CLUSTERING answers what is the boundary of the cluster that includes that point

CLASSIFICATION answers what color the point should be

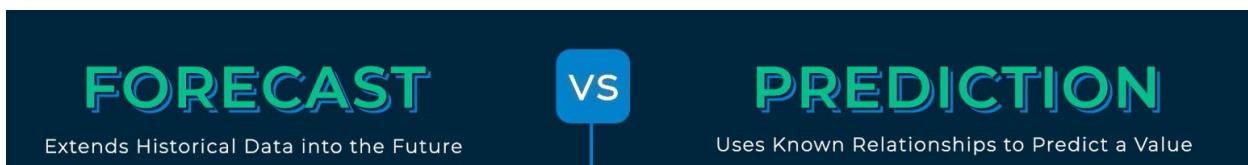
CLUSTERING is unsupervised

- Does *not* require the training dataset to contain known group values

CLASSIFICATION is supervised

- Requires the training data to contain known groups to assign to datapoints

63. Time Series Forecasting



The diagram illustrates the relationship between two data tables. On the left, a table shows 'AUTOCORRELATION' with data for 'TIME' and 'HEIGHT' at various lags (0, 1, 2). On the right, a table shows 'STANDARD CORRELATIONS' with data for 'HEIGHT', 'P1', 'P2', 'P3', and 'TIME FROM TAKEOFF'. Colored arrows connect corresponding columns between the two tables, indicating how the autocorrelation at different lags relates to the standard correlations.

TIME	HEIGHT
00:00:00	0
00:02:00	60
00:04:00	190
00:06:00	315
00:08:00	?

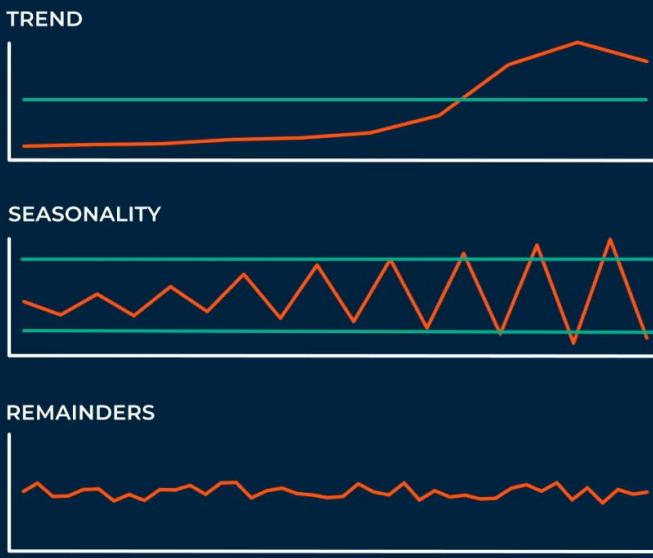
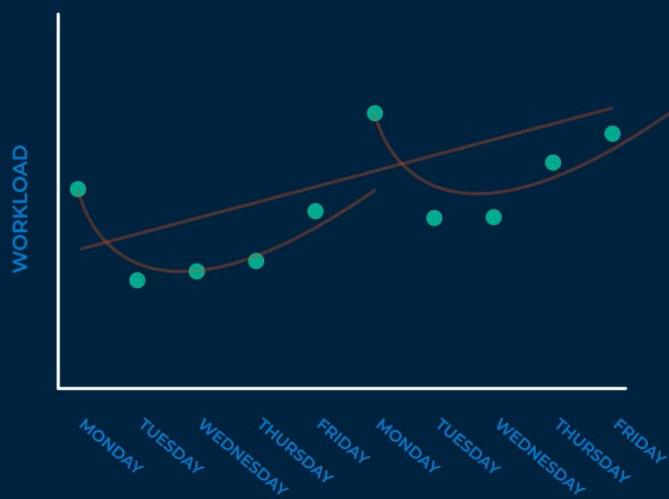
AUTOCORRELATION
LAG 0, LAG 1, LAG 2

HEIGHT	P1	P2	P3	TIME FROM TAKEOFF
1820	12	36	1.62	1.62
1934	10	36	1.57	1.57
?	?	?	?	Future

STANDARD CORRELATIONS

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

DECONSTRUCTION



Stationary data has no trend. The **trend** would be equal to the **mean** of the data.

Seasonality is “amplifying” over time, making it **non-stationary**

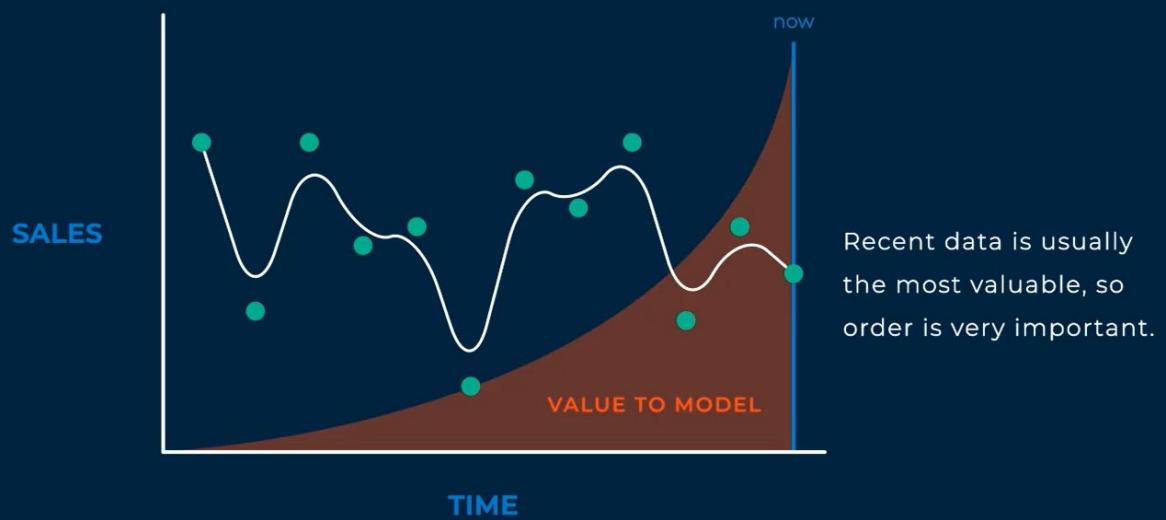
STATIONARITY

REGULARLY SPACED INTERVALS

	DATE	QUANTITY	UNIT
Irregular Month	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

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

SOME DATA IS MORE VALUABLE



TIME SERIES MODELS

- Short life span
- Projecting further into the future reduces certainty
- Model assumptions change with each new data point

PREDICTIVE MODELS

- Potentially infinite life span
- Unclear relationships between variables reduces certainty
- 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

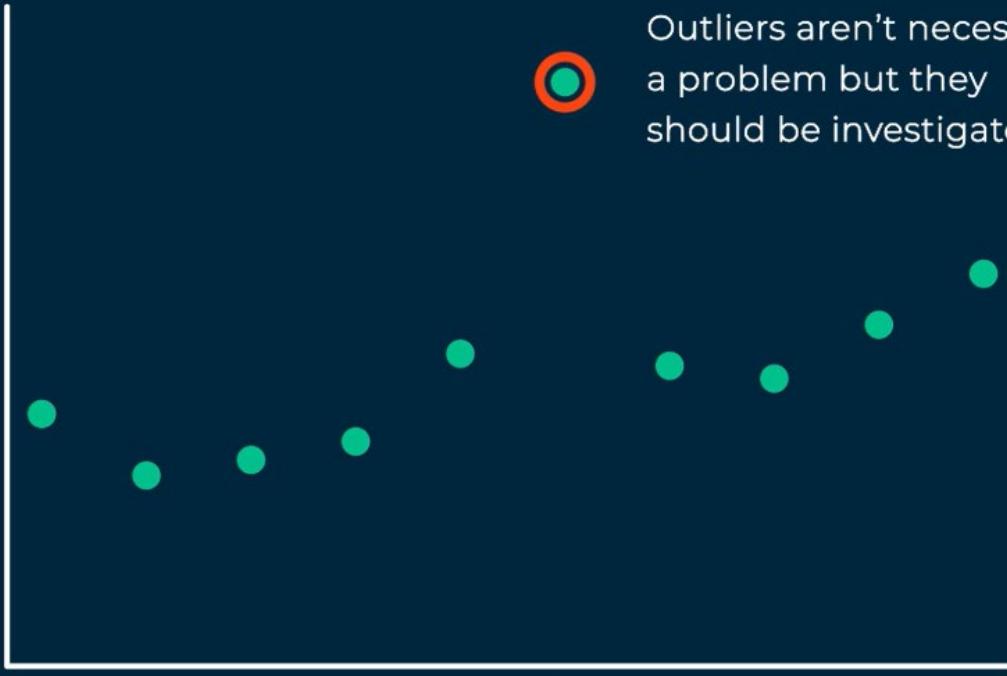
► target variable

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

OUTLIERS



Outliers aren't necessarily
a problem but they
should be investigated



An example of capping outliers:

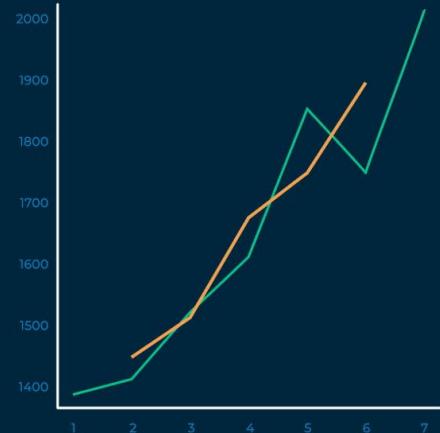
OUTLIERS



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

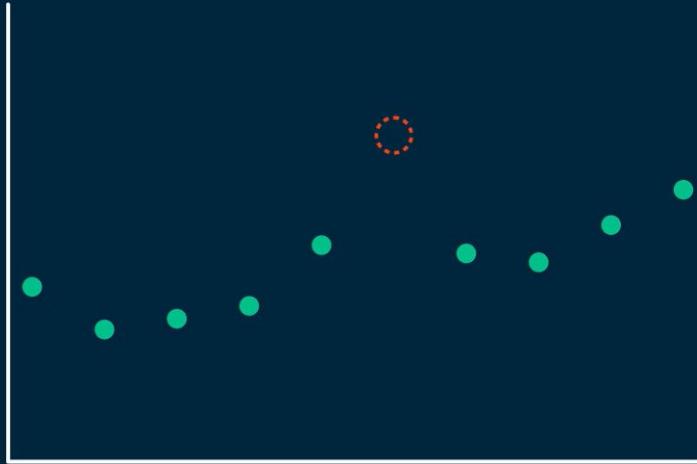
WINDOWING

DATE	QUANTITY	WINDOW
✓ 2014-05-10	1372	
✓ 2015-05-10	1406	1434.33
✓ 2016-05-10	1525	1521.66
✗ 2017-06-10	1634	1683.66
✓ 2018-05-10	1892	1758.66
✓ 2019-05-10	1750	1893.66
✓ 2020-05-10	2039	



MISSING VALUES

If you have missing values, they potentially impact the **TREND**, **SEASONALITY**, and **RESIDUALS**.



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

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

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.ymd* × +

Results - TS Filler (3) - Unfilled Input

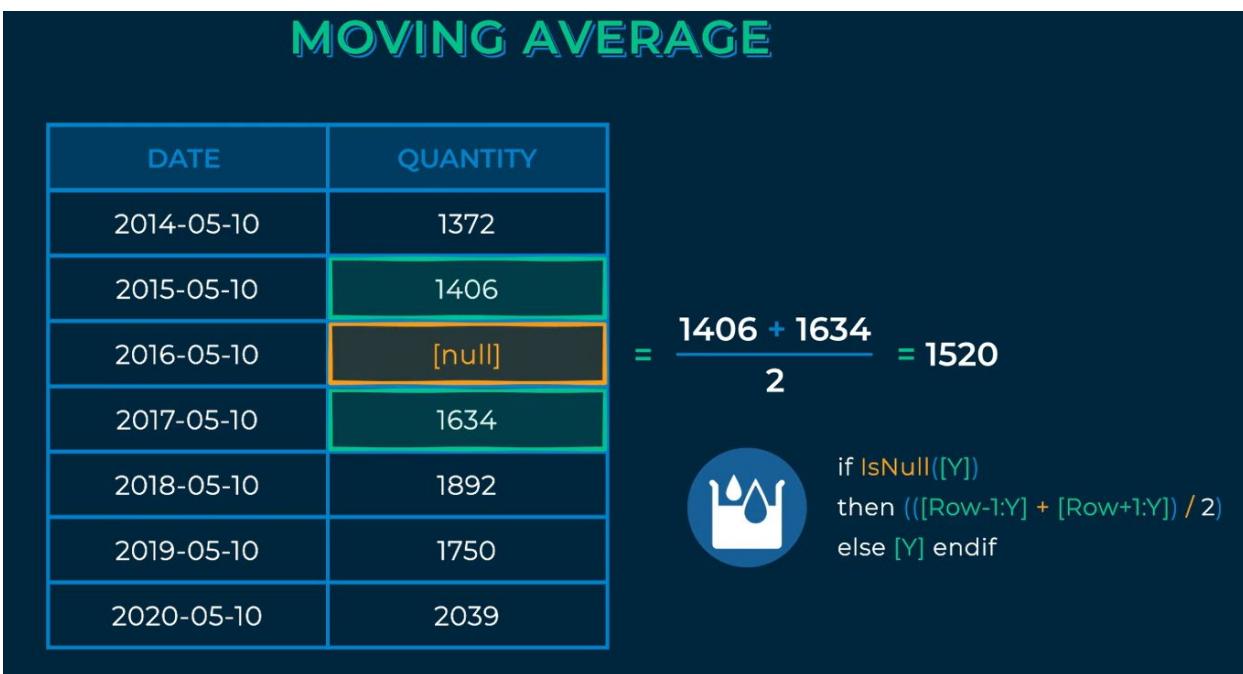
Record	DateTime_Out	Year	Month	Bookings
1	2005-06-01	2005	Jun	2138
2	2005-07-01	2005	Jul	2864
3	2005-08-01	2005	Aug	3216
4	2005-09-01	2005	Sep	1927
5	2005-10-01	2005	Oct	1415
6	2005-11-01	2005	Nov	1371
7	2005-12-01	2005	Dec	2629
8	2006-01-01	2006	Jan	3392
9	2006-02-01	2006	Feb	3246
10	2006-04-01	2006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

Results - TS Filler (3) - Filled Output

6 of 6 Fields | Cell Viewer | 79 records displayed | ↑ ↓ | Q 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

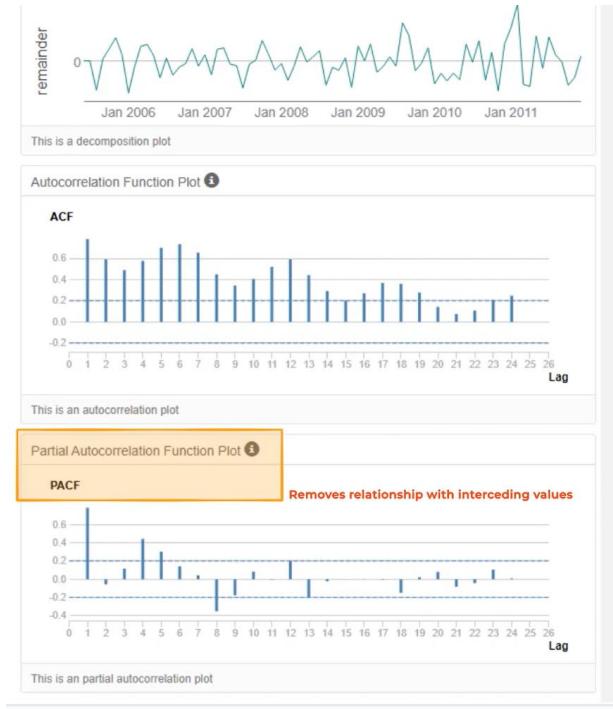
Series starting period (valid only for Target field frequency selection of Weekly, Monthly, or Annually)
The year the series starts
2005
The week, month (numeric), or quarter of the series start
6

Plot type
 Time series plot
 Seasonal plot
 Seasonal deviation plot
 Autoregression function plot
 Partial autoregression function plot
 Time series decomposition plot

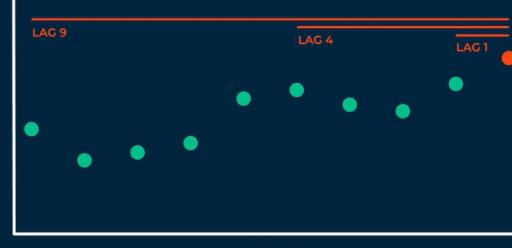
TS2.ymd*

Results - TS Plot (8) - Input

Record	DateTime_Out	OriginalDateTime	FlagGeneratedRow	Year	Month	Bookings	Bookings1
1	2005-06-01	2005-06-01 000000	False	2005	Jun	2138	2138
2	2005-07-01	2005-07-01 000000	False	2005	Jul	2864	2864
3	2005-08-01	2005-08-01 000000	False	2005	Aug	3216	3216
4	2005-09-01	2005-09-01 000000	False	2005	Sep	1927	1927
5	2005-10-01	2005-10-01 000000	False	2005	Oct	1415	1415
6	2005-11-01	2005-11-01 000000	False	2005	Nov	1371	1371
7	2005-12-01	2005-12-01 000000	False	2005	Dec	2629	2629
8	2006-01-01	2006-01-01 000000	False	2006	Jan	3392	3392
9	2006-02-01	2006-02-01 000000	False	2006	Feb	3246	3246
10	2006-03-01	[Null]	True	[Null]	[Null]	[Null]	2485
11	2006-04-01	2006-04-01 000000	False	2006	Apr	1723	1723
12	2006-05-01	[Null]	True	[Null]	[Null]	[Null]	2216
13	2006-06-01	2006-06-01 000000	False	2006	Jun	2709	2709



AUTOCORRELATION & PARTIAL AUTOCORRELATION



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
- Capping the values
- Reduce the size of the confidence intervals
- Replace with the Moving Average

Question 2

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

- Weeks
- Years
- Days
- Months
- Quarters

Question 3

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

- A prediction for the next day
- A prediction for the next week
- A prediction for the next month

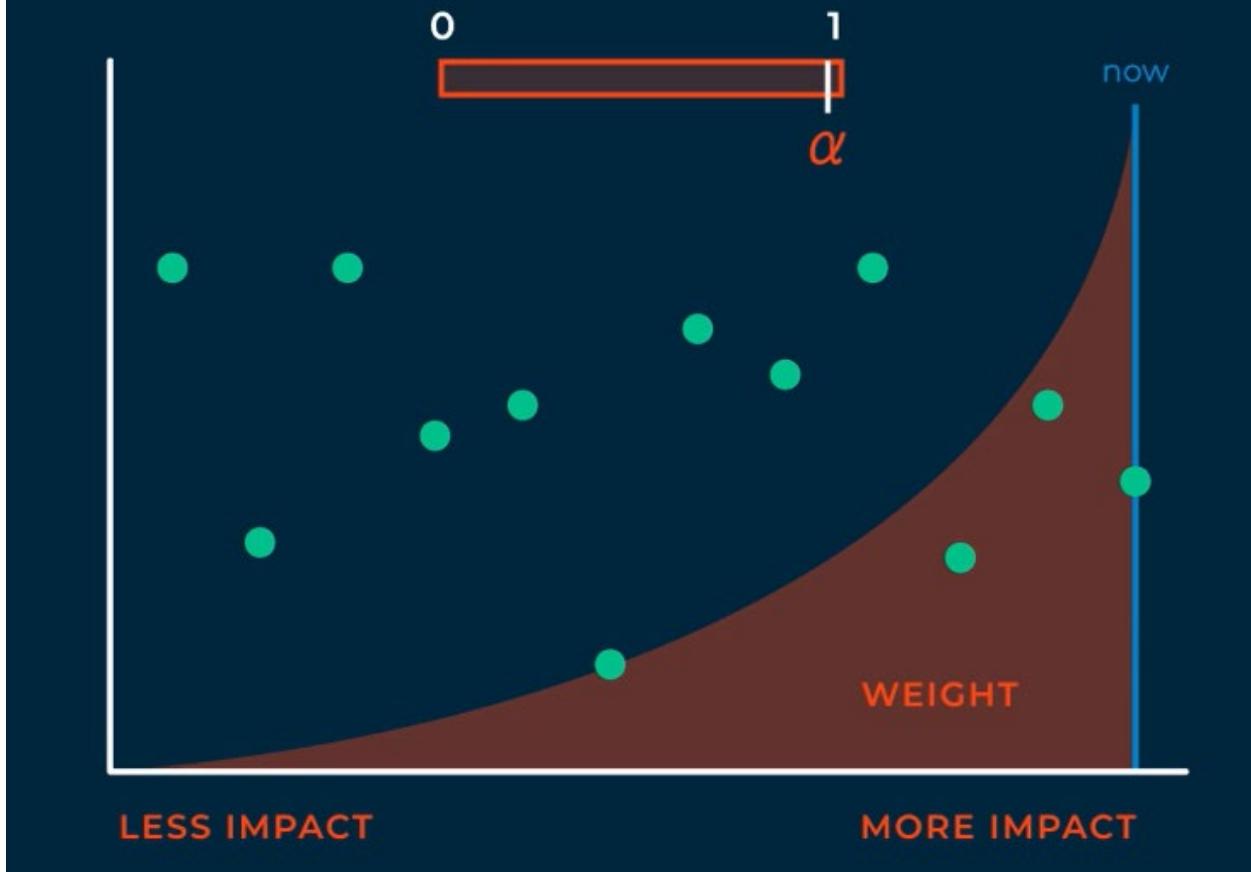
Year	Week_Of	Total Consumption EUR	Total Price EUR	Total Consumption SIB	Total Price SIB
1	2006	2006-09-01 00:00:00	11460707	58252.76	3146358
2	2006	2006-09-08 00:00:00	11882622	78173.55	3171588
3	2006	2006-09-15 00:00:00	12333684	80129.9	3178116
4	2006	2006-09-22 00:00:00	12485116	80867.01	3395715
5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

64. ETS & ARIMA



ETS

Error Trend Seasonality

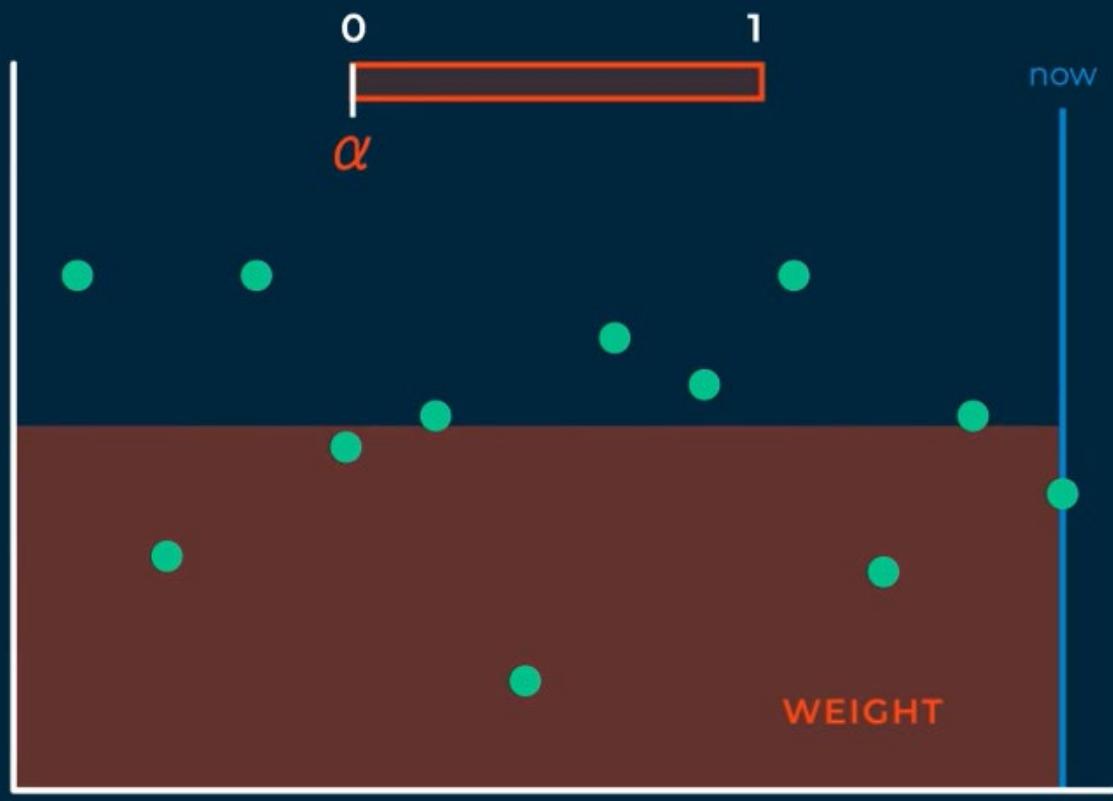


With ETS, if alpha is set to 0, then all data points are weighted equally.



ETS

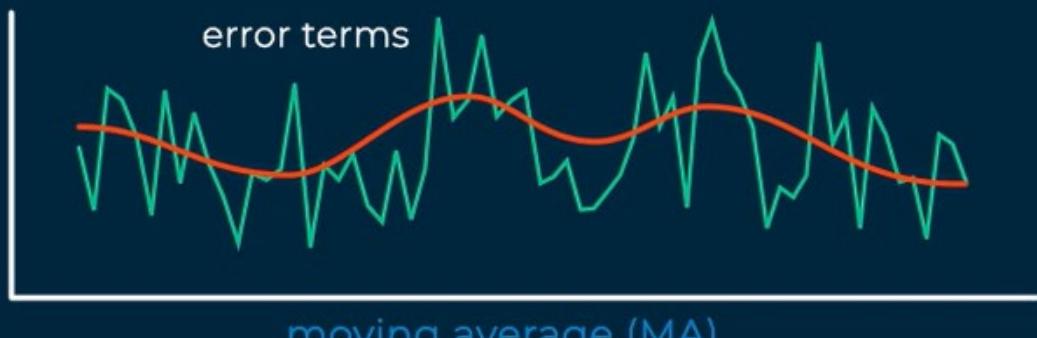
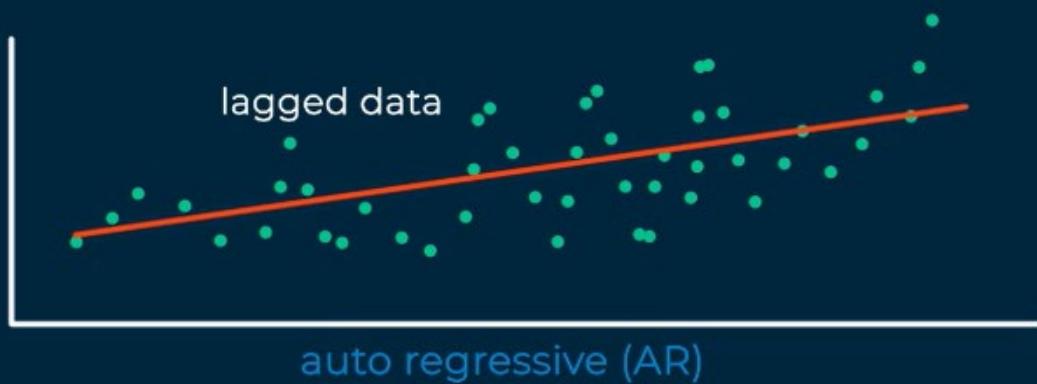
Error Trend Seasonality





ARIMA

*AutoRegressive
Integrated Moving Average*



INTEGRATED

Level of Differencing

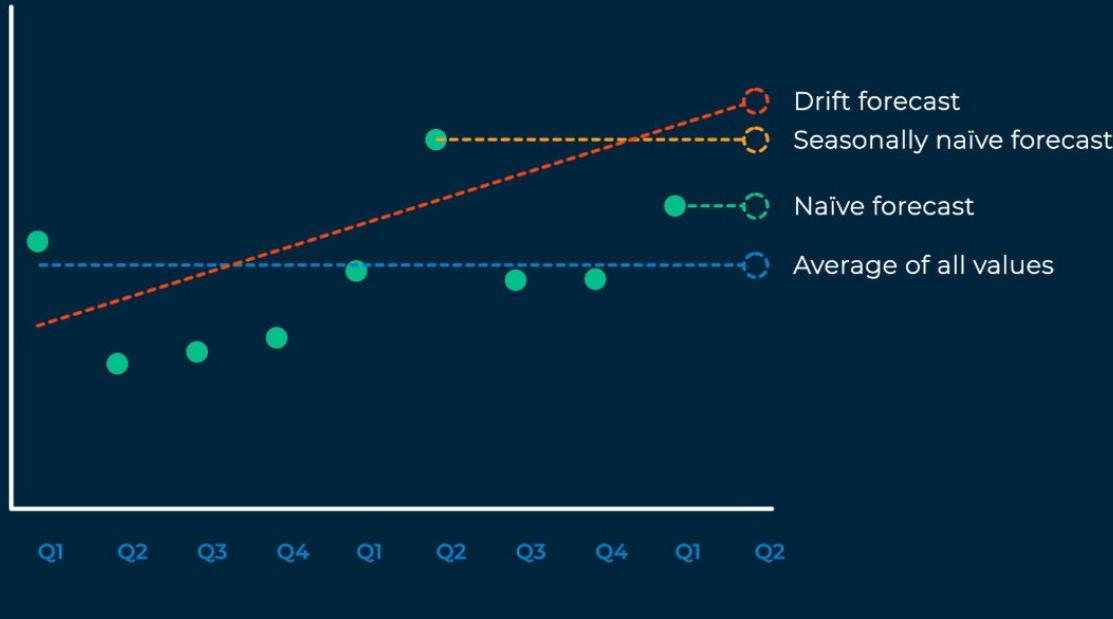
DIFFERENCING

Makes Data Stationary

AUTOCORRELATION

Similarity of Observations
at Various Lags

SIMPLE FORECASTING METHODS



STATIONARITY

- Many time series algorithms assume a dataset's mean and variance are stable over time.
- Stationary data reduces forecast uncertainty.
- Most real world data is non-stationary and will require mathematical transformation.

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.



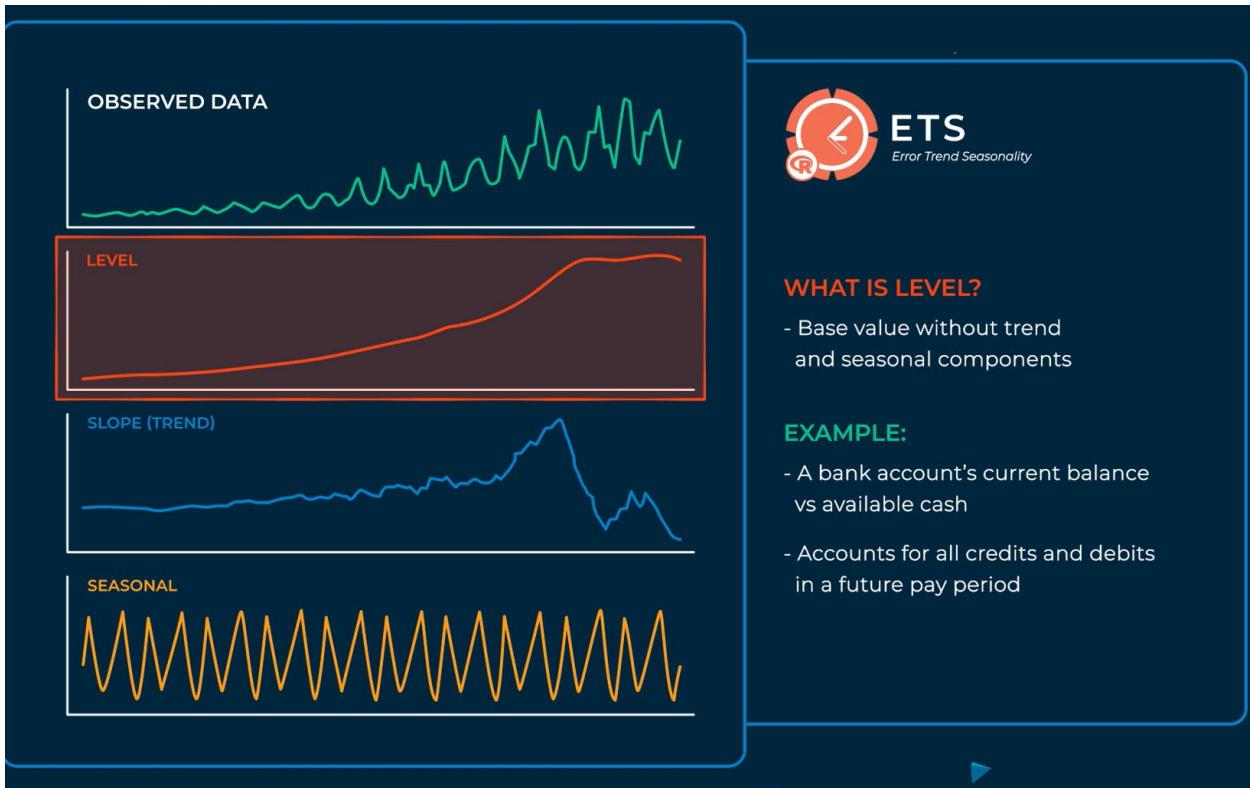
**BOX-COX
TRANSFORMATION**



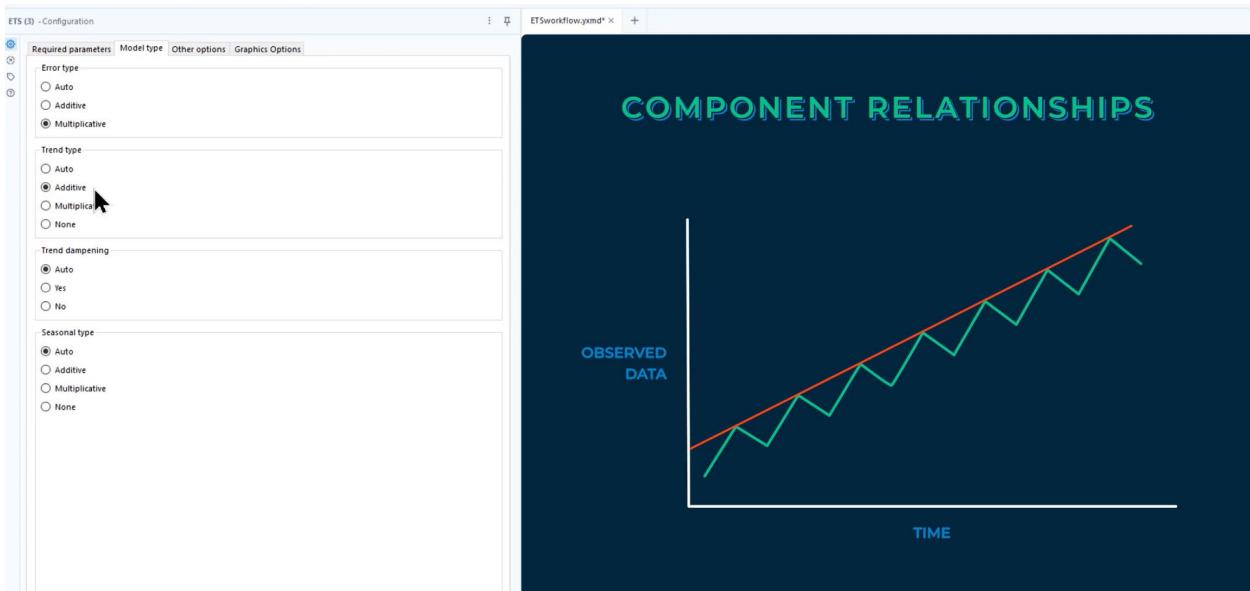
BOX-COX TRANSFORMATION

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

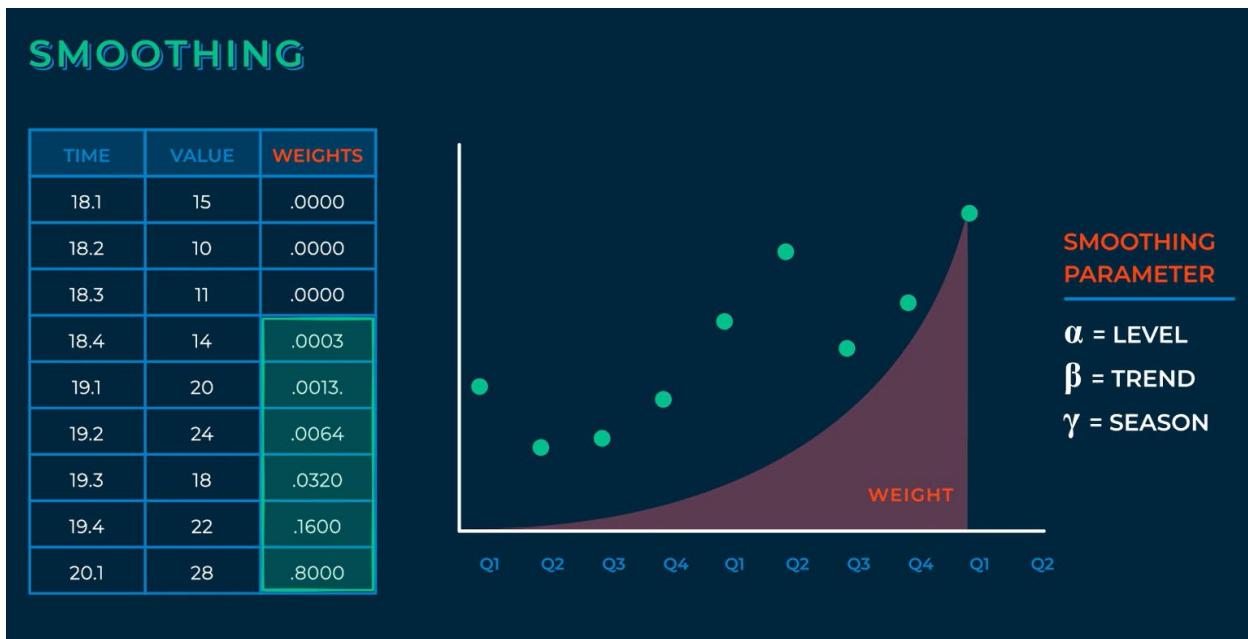
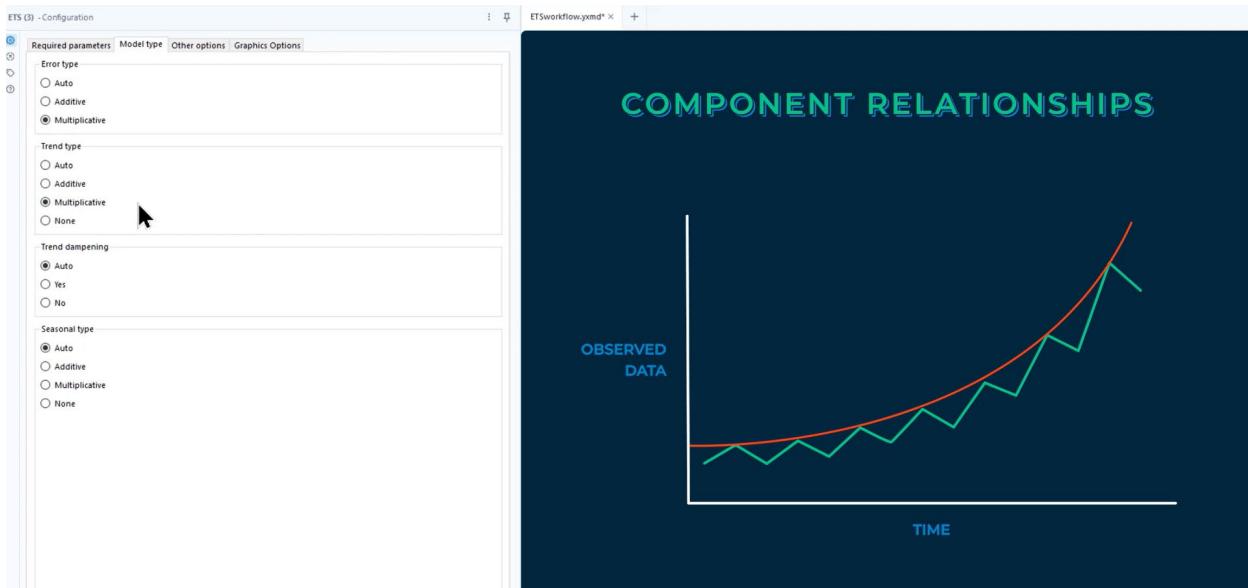
TS algorithms will adjust mean
stabilization automatically.



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.

ϕ = DAMPENING

Report Profile

1 of 1 Fields | Records 1 to 9 |

Record Report

Plots of Time Series Exponential Smoothing Model ETSEExample

In statistics, a time series is a sequence of data points measured at successive points in time. Examples of time series are the daily closing value of a stock market index or the annual flow. series analysis comprises methods for analyzing time series data in order to extract meaning characteristics of the data.

2

Decomposition by ETS(A,A,A) method

The figure displays four stacked time series plots from 2002 to 2008. The top plot shows the 'observed' data with a horizontal baseline at approximately 8.0. The second plot shows the 'level' component, which is a smooth, upward-sloping line. The third plot shows the 'slope' component, which fluctuates around zero. The bottom plot shows the 'season' component, which exhibits a clear seasonal pattern with peaks every year.

Methods used for Error, Trend, and Seasonality, respectively.

In this example, each used "Additive."

components.
Decomposition method is often used to yield information about time series components i.e. trend, cycle, seasonal, etc.

- Observed: This is the actual data.
- Level: This is the overall baseline without seasonal trends.
- Slope: This is the rate of change associated with the Level.
- Season: This shows the seasonal trend of the data.

Not all of the above components will

Trend may include a subscript to indicate dampening.

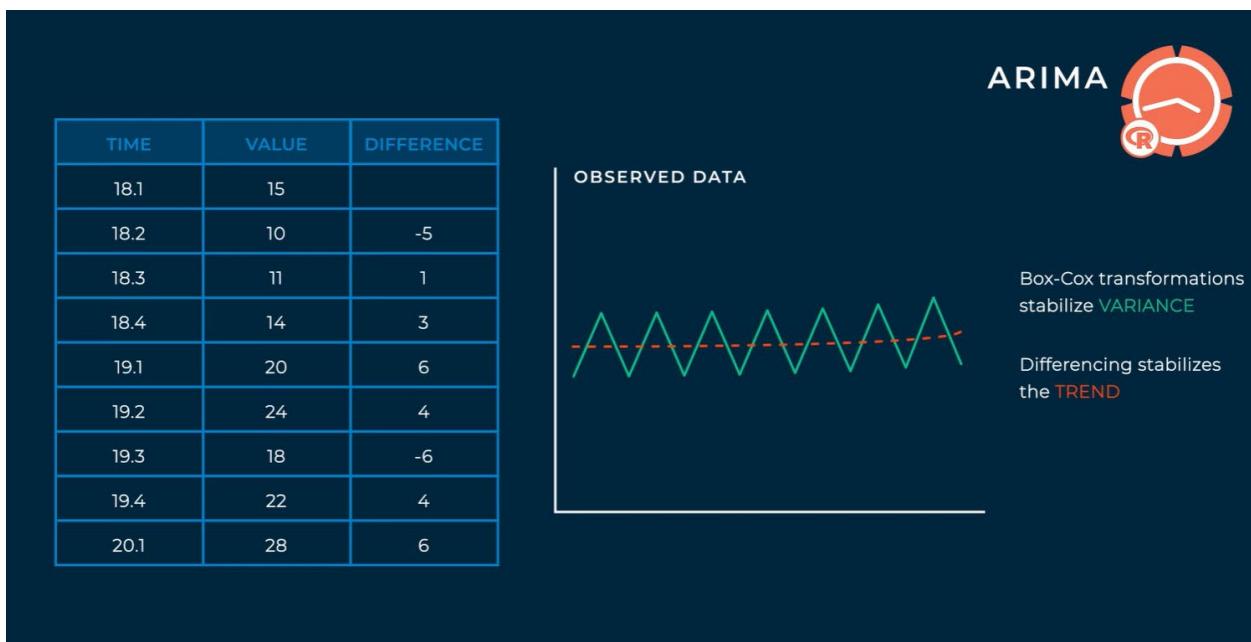
A = additive 

M = multiplicative

N = none

d = dampening

ARIMA



ARIMA

p d q

ALL PARAMETERS

0 = not present

d PARAMETER

1 = differenced once
2 = differenced twice

p, q PARAMETERS

1 = one coefficient
2 = two coefficients

Summary of ARIMA Model

Method: ARIMA(0,1,1)(1,1,0)[12]

Call:
auto.arima(Bookings)

Coefficients:

	ma1	sar1
Value	-0.671117	-0.449177
Std Err	0.077917	0.099353

σ^2 estimated as 1578141.83851: log likelihood = -916.01114

Information Criteria:

AIC	AICc	BIC
1838.0223	1838.2553	1846.0408

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-14.6510224	1175.1059989	797.4171602	-3.2161678	16.0128095	0.5634056	-0.0071189

Ljung-Box test of the model residuals:
Chi-squared = 55.5596, df = 22, p-value = 9.9e-05

Seasonal components used capitalized P, D, and Q.

ARIMA

P D Q

ALL PARAMETERS

0 = not present

d PARAMETER

1 = differenced once
2 = differenced twice

p, q PARAMETERS

1 = one coefficient
2 = two coefficients

Summary of ARIMA Model

Method: ARIMA(0,1,1)(1,1,0)[12]
seasonal components

Call:
auto.arima(Bookings)

Coefficients:

	ma1	sar1
Value	-0.671117	-0.449177
Std Err	0.077917	0.099353

σ^2 estimated as 1578141.83851: log likelihood = -916.01114

Information Criteria:

AIC	AICc	BIC
1838.0223	1838.2553	1846.0408

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-14.6510224	1175.1059989	797.4171602	-3.2161678	16.0128095	0.5634056	-0.0071189

Ljung-Box test of the model residuals:
Chi-squared = 55.5596, df = 22, p-value = 9.9e-05

Periods in the seasonal cycle:

Method: ARIMA(0,1,1)(1,1,0)[12]

seasonal components

The screenshot shows the ARIMA (11) - Configuration dialog box on the left and a COVARIATES sidebar on the right.

ARIMA (11) - Configuration

- Required parameters: Model name (ARIMAXample), Select the target field (Bookings), Use covariates in model estimation? (Optional) (checkbox checked, showing Year, Month, Bookings), Base forecast values on (Mean of covariates selected), Target field frequency (Monthly selected).
- Model customization (optional): None selected.
- Other options: Graphics Options.

COVARIATES

- Predictor Variables**: One-hot encoded for categorical data, e.g. A known holiday or annual sale, Must be projected into the future.
- Leading Indicators**: Do not need to be forecast, Precede a change in target variable, e.g. A decline in interest rate today leading to more loan applications next month.
- Must be correlated with "noise"

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.

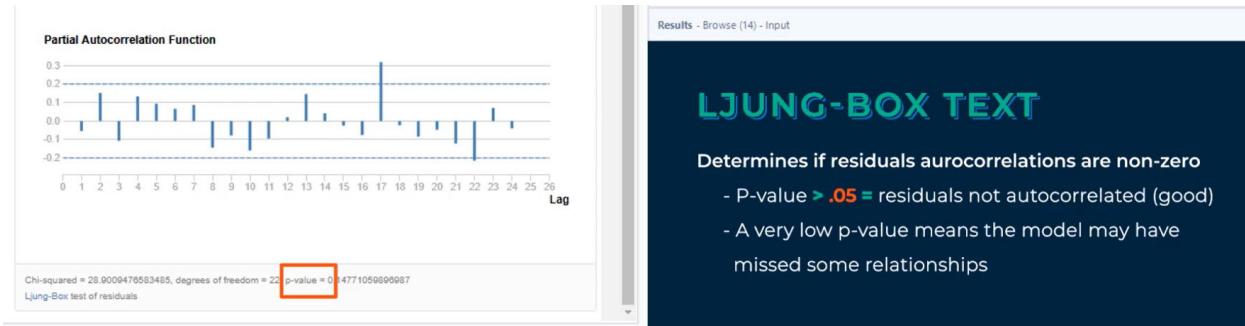
Do full enumeration of models (slow) instead of stepwise selection (faster)

The maximum order of the model (i.e. the maximum allowed value of $p + q + P + Q$)

Use multiple cores (if possible)

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

Allow drift



Question 1

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

- ARIMA with Covariates
- None of these
- ETS
- 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
- None

This is neither Linear nor Exponential



Question 4

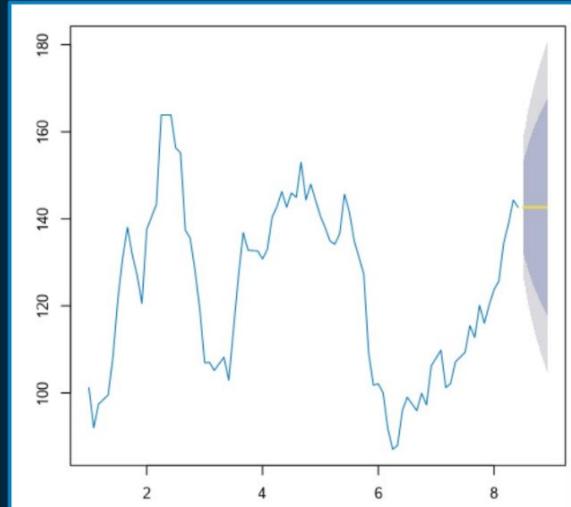
What is the purpose of a Box-Cox transformation?

- 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?

- Damped 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



TIME	VALUE
00:01:00	383
00:02:00	366
00:03:00	250
00:04:00	318
00:05:00	334
00:06:00	397
00:07:00	575
00:08:00	701
00:09:00	506
00:10:00	335
00:11:00	372

PREPPING DATA FOR EVALUATION

- Split **TRAINING** and **VALIDATION** data
- Validation set must be about 20% of most recent data
- If 20% isn't possible, at least as many values as you plan to forecast
- A Filter Tool accomplishes this easily

Browse (50) - Configuration

Report Profile | 4 records displayed, 2 fields, 59 KB |

1 of 1 Fields | Records 1 to 4 |

Record Report

1 Comparison of Time Series Models

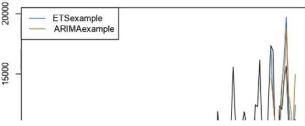
2 Actual and Forecast Values:

Actual ETSExample ARIMAxexample	
17363	172094.1931
16952	129649.0295
9012	105273.3057
7730	86740.7286
12345	10114.20556
12055	139692.2977
14496	162353.53814
15854	197353.3285
11063	12605.1342
7934	100561.9086
7106	8043.99014
11209	12404.39119

3 Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETSExample	-804.8949	2170.557	1863.474	-8.7758	15.8353	1.5274
ARIMAxexample	-1126.9956	2734.298	2479.965	-14.4114	22.8723	2.0327

4 Actual and Forecast Values



METRICS



ME	RMSE	MAE	MPE	MAPE	MASE	ACFI
18.83	880.18	510.01	-0.18	11.38	0.38	0.24



ME	RMSE	MAE	MPE	MAPE	MASE	ACFI
97.10	979.55	653.10	-1.47	15.29	0.49	-0.007

- In-Sample Statistics
- Measures accuracy of forecast against training data
- Each metric represents ERROR in a different way



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

- Out-of-Sample Statistics
- Measures accuracy of models against unseen (testing) data
- Displays ERROR metrics by model type for easy comparison

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



MODEL	ME	RMSE	MAE	Mean Absolute Error		
				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

MPE and MAPE can be compared between models.



MODEL	ME	RMSE	MAE	Mean Absolute Percentage Error		
				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

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



MODEL	ME	RMSE	MAE	Mean Absolute Scaled Error		
				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

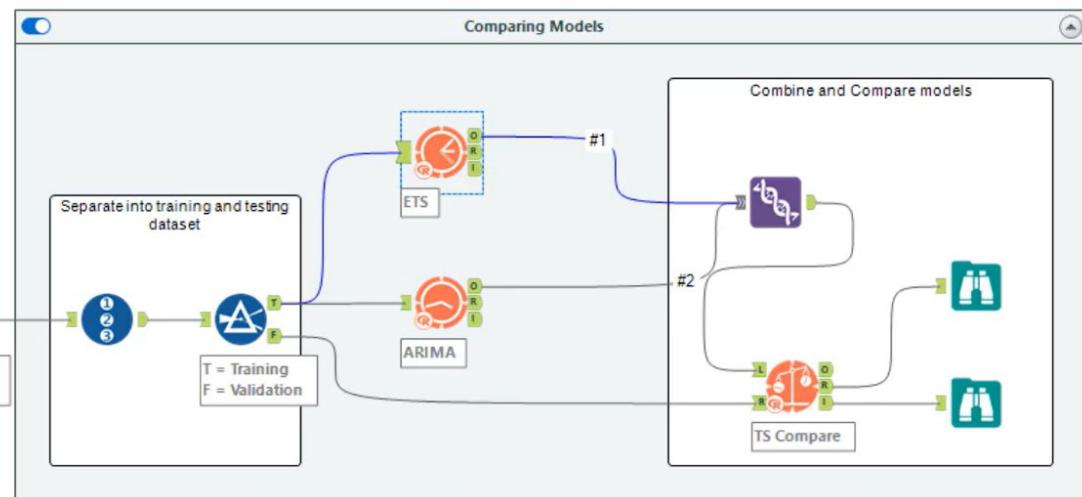
Values with the lowest absolute value indicate better performance.



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.

The screenshot shows a data processing interface with three main components:

- Text Input (22) - Configuration:** A table view showing data from 2002 to 2004. The last column, "Workers", is highlighted with a red box.
- Covariate:** A flow diagram illustrating the data pipeline. It starts with an ARIMA model, which feeds into a "TS Covariate Forecast" block. This forecast is then used to generate an "Estimated Covariate for prediction periods".
- Results - Text Input (41) - Output:** A table viewer showing the same data as the input, but with additional columns for "Record" and "Workers". The "Workers" column is highlighted with a red box. Two arrows point to specific rows: one labeled "Next time period" pointing to row 1 (value 500), and another labeled "Furthest time period" pointing to row 6 (value 539).

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.

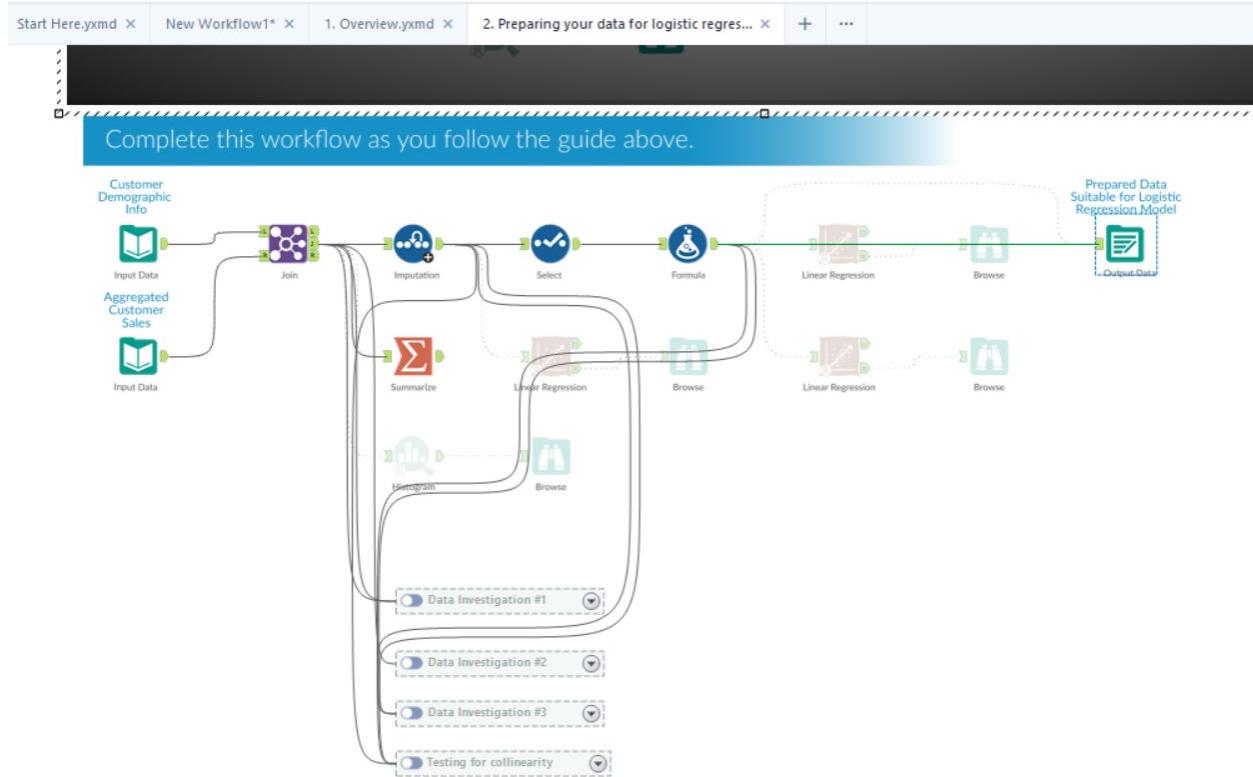
Question 3

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](#)

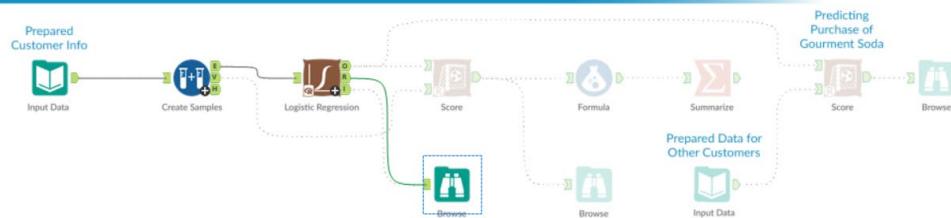
66. Examples



The Deviance Residuals section gives us information about how well our model performed on the training data. Generally, having a Deviance Residual greater than 2 or less than -2 can indicate a model that fits poorly. However, we have a large dataset, and the majority of our deviance residuals are quite small in magnitude. (Half of them are between -2.11×10^{-8} and 2.11×10^{-8} .)

	Min	1Q	Median	3Q	Max
-	-2.45×10^{-8}	-2.11×10^{-8}	-2.11×10^{-8}	2.11×10^{-8}	2.46×10^{-8}

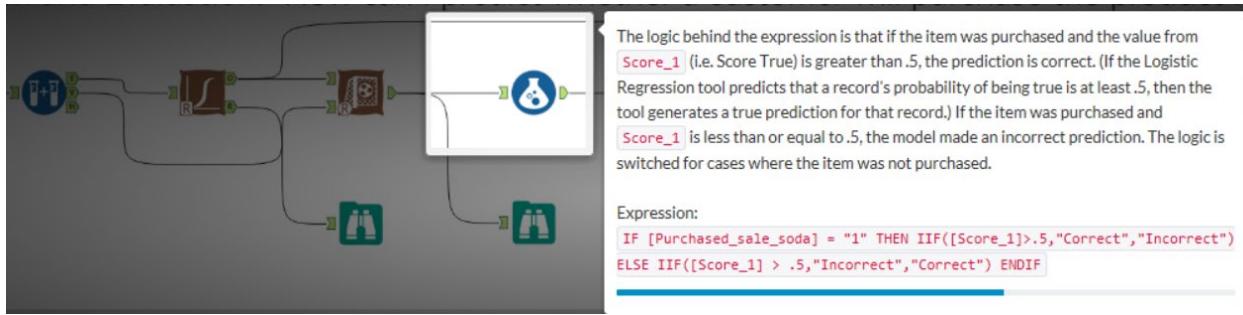
Complete this workflow as you follow the guide above.



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**



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](#)

Model Creation and Evaluation: How can I predict how much a customer spent?

To determine if a predictor is significant we must first select a confidence interval (CI). We will use a 95% CI. Now we observe the last column of the Coefficients table. The last column is marked by a ., *, **, or ***. The period is significant at the 90% CI. A single star is significant at the 95% CI. Two stars are significant at the 99% CI. Three stars are significant at the 99.9% CI. Therefore, we want variables with one or more stars because it meets our 95% CI criteria.

	2.1676690	4.327e+00	0.50091	0.6165
(Intercept)	-0.0002164	4.092e-04	-0.52876	0.59704
TimeSpentOnSite	-1.2996417	2.053e+00	-0.63298	0.52683
DeviceTypeMobile	-1.1634863	2.125e+00	-0.54753	0.58408
DeviceTypeTablet	2.8303360	3.152e+00	0.89792	0.36935
BrowserTypeFirefox	-0.5461563	1.997e+00	-0.27342	0.78456
BrowserTypeInternet Explorer	-3.7518619	4.407e+00	-0.85132	0.39471
BrowserTypeOther	-1.3933162	4.984e+00	-0.27954	0.77987
BrowserTypeSafari	-4.2345043	4.022e+00	-1.05294	0.29251
DisplayNetworkAdsterra	31.1755488	2.898e+00	10.75725	< 2.2e-16 ***
DisplayNetworkGoogle AdSense	0.1876643	3.576e+00	0.05248	0.95816
DisplayNetworkInfolinks	1.3336920	2.590e+00	0.51332	0.60779
DisplayNetworkMedia.net	-0.7363423	3.237e+00	-0.22750	0.82006
DisplayNetworkYahoo	0.1289893	1.455e+00	0.08927	0.92888
GenderMale	10.0548759	7.007e-02	143.50590	< 2.2e-16 ***
Age	0.0049626	5.885e-05	84.32936	< 2.2e-16 ***
HouseholdIncome	NA	NA	NA	NA
BrowserTypeMobile	NA	NA	NA	NA

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

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

As you can see, the other display networks are included. We will keep them because of the way the Linear Regression tool is configured. The significance of Google Adsense outweighs the insignificance of the other display networks. Therefore they will be part of our regression model. We will move forward with the identified significant variables.

The Multiple R-squared and Adjusted R-squared values are known as the coefficient of determination or goodness of fit values. In our model, they are respectively 0.9768 and 0.9767. This means that the model fits 97.7% of our data. The linear model is a strong representation of the data used. Therefore the predictor variables are highly correlated with customer spend.

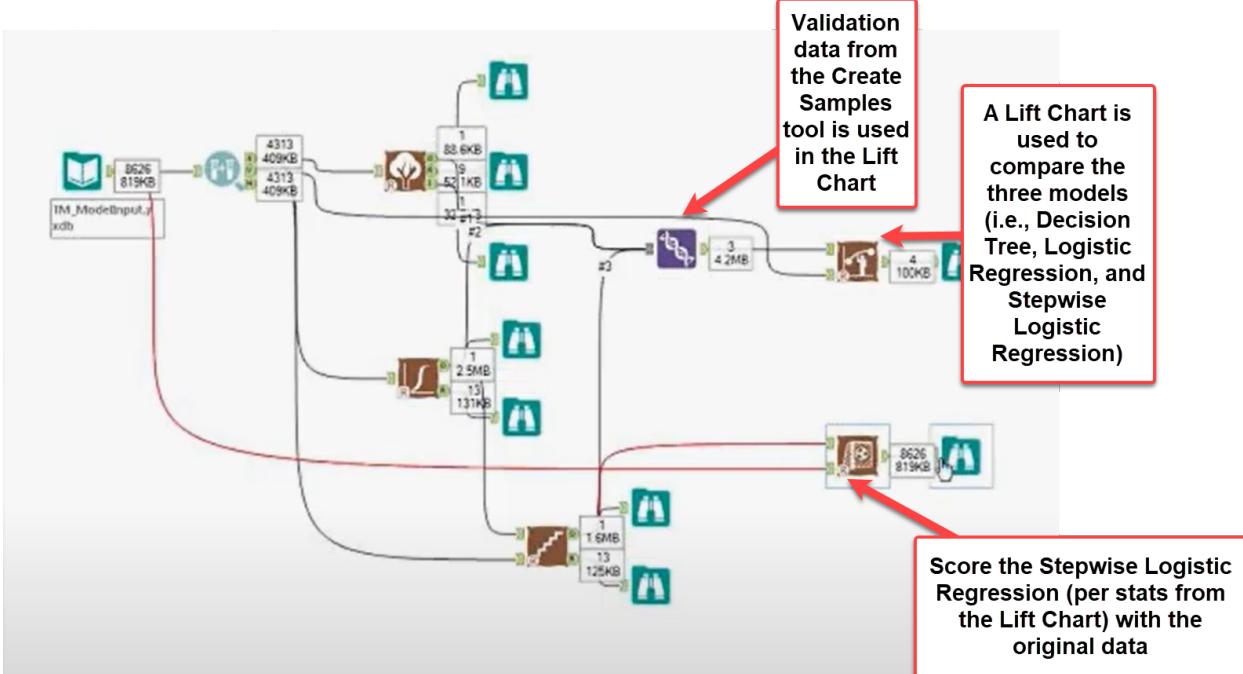
A high R-squared value is a necessary measure of model quality, but is not sufficient when used alone. Be careful when using R-squared. It can be manipulated by adding predictor variables, even if the variables are insignificant. However, the Adjusted R-squared attempts to account for this issue by penalizing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

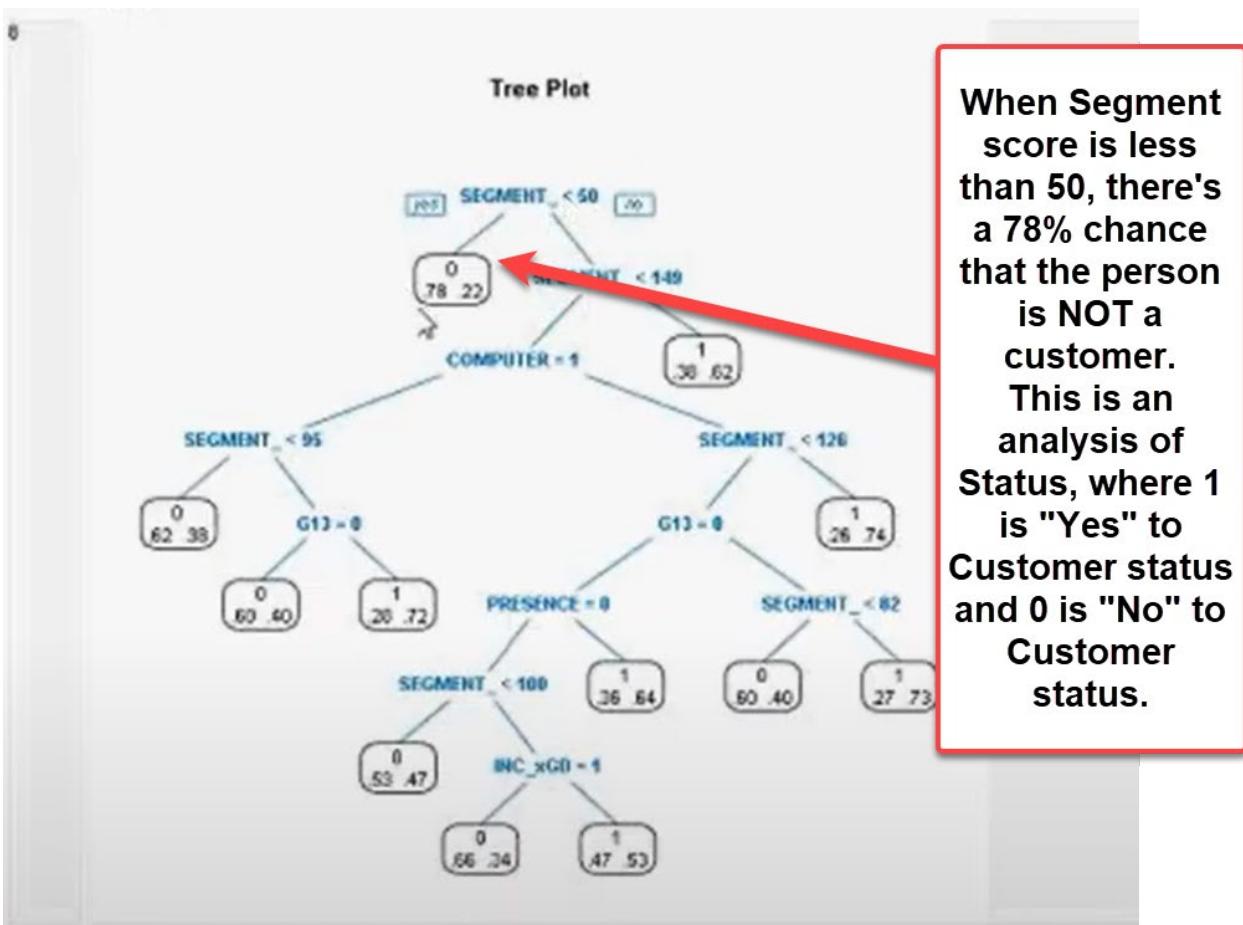
67. [Tool Mastery | Association Analysis](#)
68. [Tool Mastery | Score](#)
69. [Pre-Predictive: Using the Data Investigation Tools - Part 3 of 4](#)
70. [Box plot](#)
71. [Weekly Challenge #157: An Expert Challenge](#)
72. Linear Regressions and categorical variables

Record	Report																																																																																																				
Report for Linear Model LinMod																																																																																																					
1 Basic Summary																																																																																																					
2 Call: lm(formula = Food_Away ~ Income + AGE_REF + Ed_Attain_Ref + FAM_SIZE + Fam_Struc + O_Child + Urban, data = the.data)																																																																																																					
3 Residuals:																																																																																																					
<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th></th> <th>Min</th> <th>1Q</th> <th>Median</th> <th>3Q</th> <th>Max</th> </tr> </thead> <tbody> <tr> <td></td> <td>-6425</td> <td>-617</td> <td>-247</td> <td>318</td> <td>14620</td> </tr> </tbody> </table>			Min	1Q	Median	3Q	Max		-6425	-617	-247	318	14620																																																																																								
	Min	1Q	Median	3Q	Max																																																																																																
	-6425	-617	-247	318	14620																																																																																																
4 Coefficients:																																																																																																					
<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th></th> <th>Estimate</th> <th>Std. Error</th> <th>t value</th> <th>Pr(> t)</th> </tr> </thead> <tbody> <tr> <td>(Intercept)</td> <td>-2.612e+02</td> <td>9.000e+02</td> <td>-0.28762</td> <td>0.77367</td> </tr> <tr> <td>Income</td> <td>9.021e-03</td> <td>5.231e-04</td> <td>17.24467</td> <td>< 2.2e-16 ***</td> </tr> <tr> <td>AGE_REF</td> <td>-2.496e+00</td> <td>2.380e+00</td> <td>-1.04862</td> <td>0.29453</td> </tr> <tr> <td>Ed_Attain_Ref@LTHS</td> <td>5.770e+02</td> <td>8.770e+02</td> <td>0.65797</td> <td>0.51067</td> </tr> <tr> <td>Ed_Attain_Ref@HSGrad</td> <td>5.036e+02</td> <td>8.747e+02</td> <td>0.57573</td> <td>0.56499</td> </tr> <tr> <td>Ed_Attain_Ref@SomeCol/Assoc</td> <td>5.668e+02</td> <td>8.745e+02</td> <td>0.76246</td> <td>0.44591</td> </tr> <tr> <td>Ed_Attain_Ref@Bachelor's</td> <td>8.040e+02</td> <td>8.760e+02</td> <td>1.00503</td> <td>0.31506</td> </tr> <tr> <td>Ed_Attain_Ref@Grad/Prof</td> <td>9.214e+02</td> <td>8.779e+02</td> <td>1.04961</td> <td>0.29408</td> </tr> <tr> <td>FAM_SIZE</td> <td>-3.091e+01</td> <td>4.286e+01</td> <td>-0.72126</td> <td>0.47087</td> </tr> <tr> <td>Fam_Struc@C w/Chldm</td> <td>2.146e+02</td> <td>1.361e+02</td> <td>1.57668</td> <td>0.11509</td> </tr> <tr> <td>Fam_Struc@C w/Other</td> <td>-9.399e+01</td> <td>1.099e+02</td> <td>-0.49409</td> <td>0.62075</td> </tr> <tr> <td>Fam_Struc@Other</td> <td>7.510e+00</td> <td>1.177e+02</td> <td>0.06379</td> <td>0.94915</td> </tr> <tr> <td>Fam_Struc@Single Fem</td> <td>1.211e+02</td> <td>1.206e+02</td> <td>-0.0061</td> <td>0.03765 *</td> </tr> <tr> <td>Fam_Struc@Single Male</td> <td>1.275e+02</td> <td>1.206e+02</td> <td>-0.16846</td> <td>0.86624</td> </tr> <tr> <td>Fam_Struc@Single Par</td> <td>1.969e+02</td> <td>1.206e+02</td> <td>-0.24859</td> <td>0.80372</td> </tr> <tr> <td>O_Child@6-11</td> <td>1.582e+02</td> <td>1.25006</td> <td>0.34730</td> <td>0.72842</td> </tr> <tr> <td>O_Child@12-17</td> <td>1.450e+02</td> <td>1.25006</td> <td>0.21149</td> <td></td> </tr> <tr> <td>O_Child@No Children</td> <td>1.555e+02</td> <td>1.44406</td> <td>0.14094</td> <td></td> </tr> <tr> <td>UrbanYes</td> <td>1.439e+02</td> <td>1.75077</td> <td>0.0802 .</td> <td></td> </tr> </tbody> </table>			Estimate	Std. Error	t value	Pr(> t)	(Intercept)	-2.612e+02	9.000e+02	-0.28762	0.77367	Income	9.021e-03	5.231e-04	17.24467	< 2.2e-16 ***	AGE_REF	-2.496e+00	2.380e+00	-1.04862	0.29453	Ed_Attain_Ref@LTHS	5.770e+02	8.770e+02	0.65797	0.51067	Ed_Attain_Ref@HSGrad	5.036e+02	8.747e+02	0.57573	0.56499	Ed_Attain_Ref@SomeCol/Assoc	5.668e+02	8.745e+02	0.76246	0.44591	Ed_Attain_Ref@Bachelor's	8.040e+02	8.760e+02	1.00503	0.31506	Ed_Attain_Ref@Grad/Prof	9.214e+02	8.779e+02	1.04961	0.29408	FAM_SIZE	-3.091e+01	4.286e+01	-0.72126	0.47087	Fam_Struc@C w/Chldm	2.146e+02	1.361e+02	1.57668	0.11509	Fam_Struc@C w/Other	-9.399e+01	1.099e+02	-0.49409	0.62075	Fam_Struc@Other	7.510e+00	1.177e+02	0.06379	0.94915	Fam_Struc@Single Fem	1.211e+02	1.206e+02	-0.0061	0.03765 *	Fam_Struc@Single Male	1.275e+02	1.206e+02	-0.16846	0.86624	Fam_Struc@Single Par	1.969e+02	1.206e+02	-0.24859	0.80372	O_Child@6-11	1.582e+02	1.25006	0.34730	0.72842	O_Child@12-17	1.450e+02	1.25006	0.21149		O_Child@No Children	1.555e+02	1.44406	0.14094		UrbanYes	1.439e+02	1.75077	0.0802 .	
	Estimate	Std. Error	t value	Pr(> t)																																																																																																	
(Intercept)	-2.612e+02	9.000e+02	-0.28762	0.77367																																																																																																	
Income	9.021e-03	5.231e-04	17.24467	< 2.2e-16 ***																																																																																																	
AGE_REF	-2.496e+00	2.380e+00	-1.04862	0.29453																																																																																																	
Ed_Attain_Ref@LTHS	5.770e+02	8.770e+02	0.65797	0.51067																																																																																																	
Ed_Attain_Ref@HSGrad	5.036e+02	8.747e+02	0.57573	0.56499																																																																																																	
Ed_Attain_Ref@SomeCol/Assoc	5.668e+02	8.745e+02	0.76246	0.44591																																																																																																	
Ed_Attain_Ref@Bachelor's	8.040e+02	8.760e+02	1.00503	0.31506																																																																																																	
Ed_Attain_Ref@Grad/Prof	9.214e+02	8.779e+02	1.04961	0.29408																																																																																																	
FAM_SIZE	-3.091e+01	4.286e+01	-0.72126	0.47087																																																																																																	
Fam_Struc@C w/Chldm	2.146e+02	1.361e+02	1.57668	0.11509																																																																																																	
Fam_Struc@C w/Other	-9.399e+01	1.099e+02	-0.49409	0.62075																																																																																																	
Fam_Struc@Other	7.510e+00	1.177e+02	0.06379	0.94915																																																																																																	
Fam_Struc@Single Fem	1.211e+02	1.206e+02	-0.0061	0.03765 *																																																																																																	
Fam_Struc@Single Male	1.275e+02	1.206e+02	-0.16846	0.86624																																																																																																	
Fam_Struc@Single Par	1.969e+02	1.206e+02	-0.24859	0.80372																																																																																																	
O_Child@6-11	1.582e+02	1.25006	0.34730	0.72842																																																																																																	
O_Child@12-17	1.450e+02	1.25006	0.21149																																																																																																		
O_Child@No Children	1.555e+02	1.44406	0.14094																																																																																																		
UrbanYes	1.439e+02	1.75077	0.0802 .																																																																																																		
5 Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'																																																																																																					
6 Residual standard error: 1229.2 on 1407 degrees of freedom																																																																																																					
Multiple R-squared: 0.2788, Adjusted R-Squared: 0.26																																																																																																					
F-statistic: 30.21 on 18 and 1407 DF, p-value: < 2.2e-16																																																																																																					
7 Type II ANOVA Analysis																																																																																																					
8 Response: Food_Away																																																																																																					

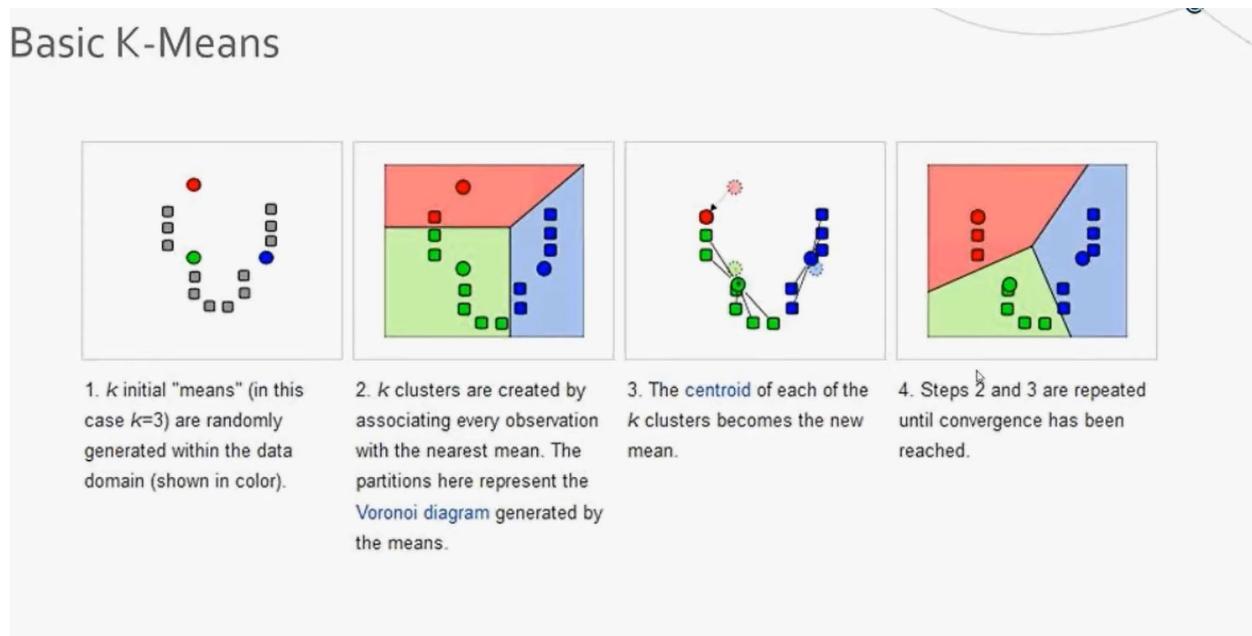
73. Example of a classification problem, from [Regression Modeling](#):



Decision Tree:



74. Basic K-Means:



75. [Cluster Analysis Basics](#)

76. Data Investigation/Prep (Key Terms):

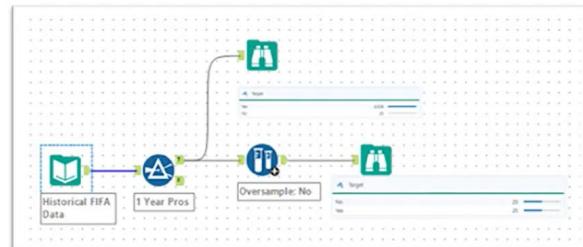
Data Investigation/Prep – Key Terms

- 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.



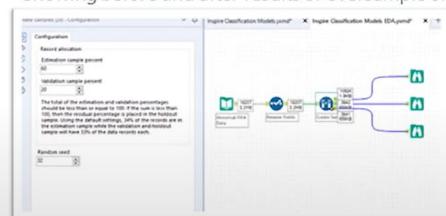
Data Investigation/Prep – Key Terms, Cont.

- 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

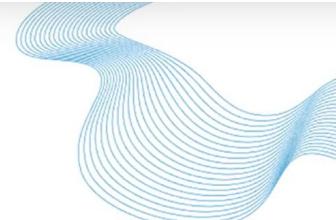


Create Sample:

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.

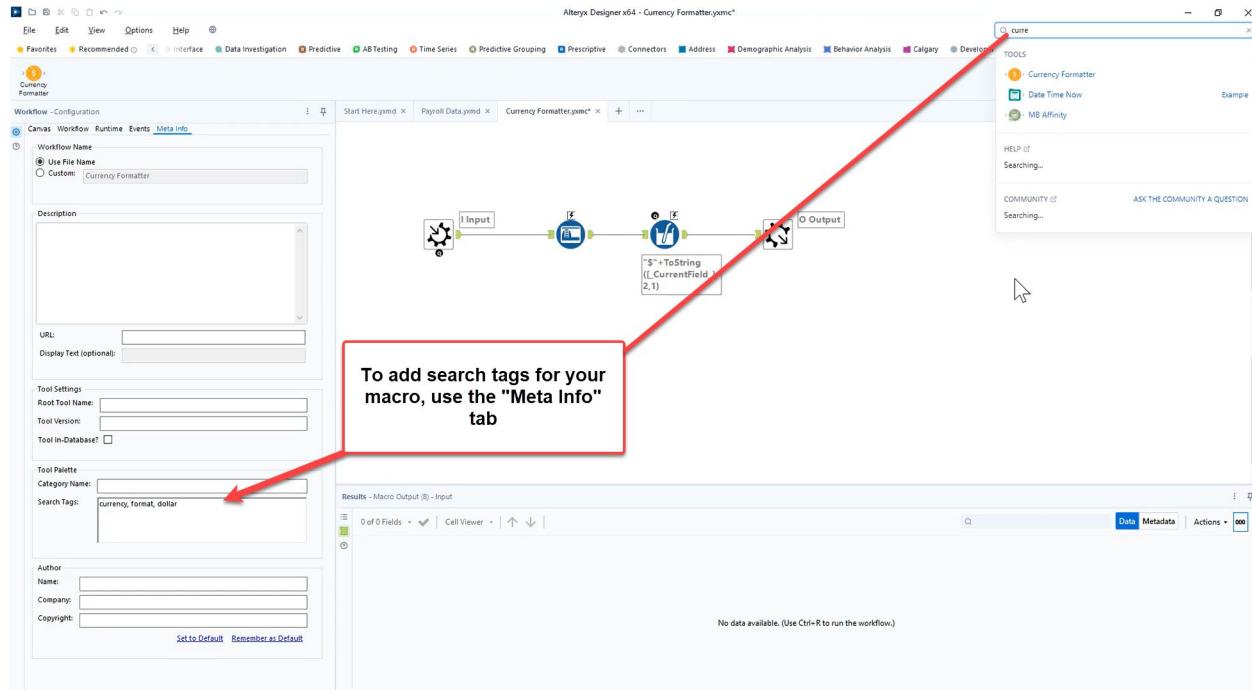
78. [Predictive Modeling: Classification](#)

79.

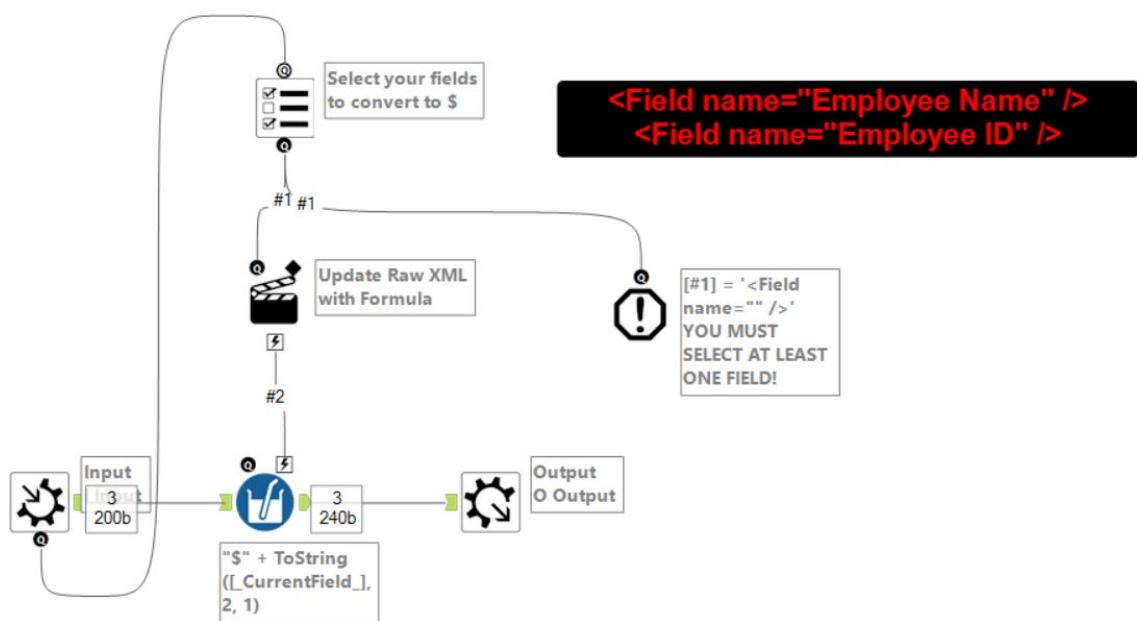
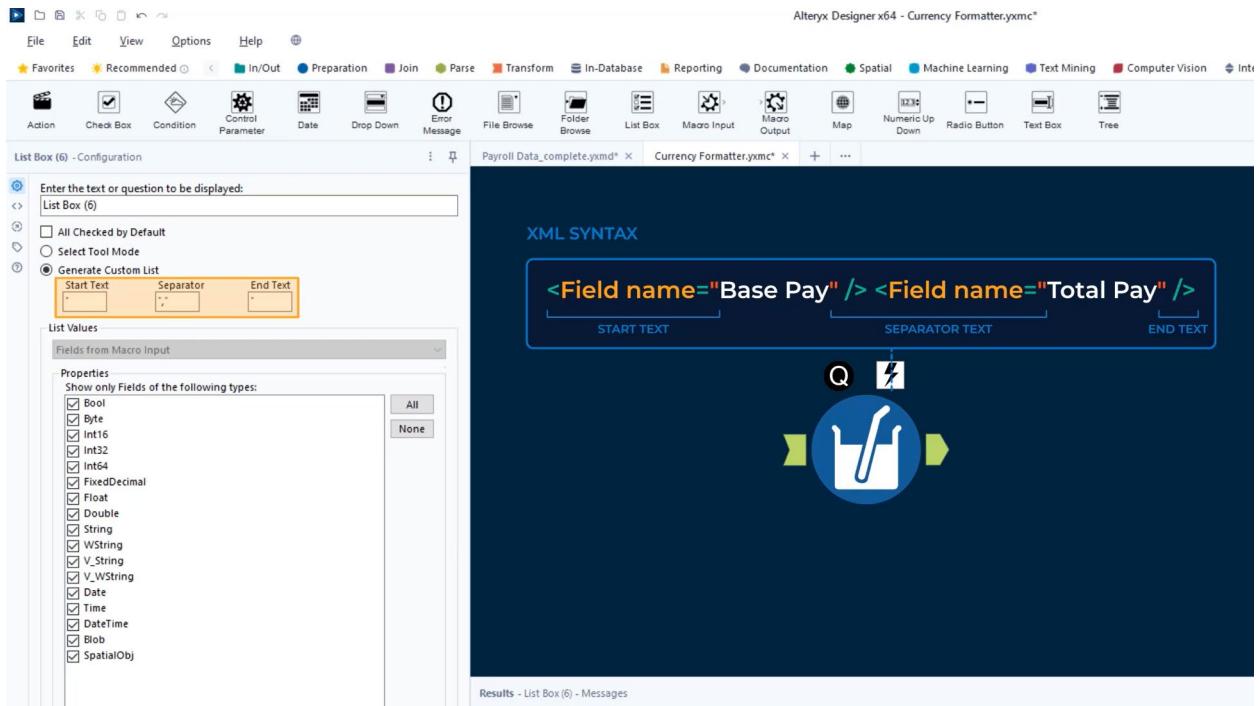
Analytic Applications & Macros

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:

The screenshot shows the Alteryx Designer interface. On the left, the 'Tax Rate Macro (17) - Configuration' pane displays two 'Group By' sections: 'Control GroupBy Field' set to 'City' and 'Input5 GroupBy Field' also set to 'City'. The main workspace on the right is titled 'CONSIDERATIONS' with the following bullet points:

- Input records **must** match with a batch group to be included in output
Unlike a Join, where unmatched records are kept in L/R anchors
- To troubleshoot, know how many records to expect in your output

Below this, the 'Results - Tax Rate Macro (17) - Output4' pane shows a table with 10 records:

Record	Order Number	Price	Order Date	City	State	Total Cost
1	05	94.74	2004-11-02	Allentown	PA	100.42
2	03	95.66	2003-10-28	Burlingame	CA	105.84
3	01	95.7	2003-02-24	NYC	NY	104.07
4	07	99.91	2003-02-24	NYC	NY	108.65
5	10	48.05	2004-11-21	NYC	NY	52.25
6	06	76.36	2005-03-03	New Bedford	MA	81.13
7	08	44.51	2005-03-03	New Bedford	MA	47.29
8	09	85.77	2005-03-03	New Bedford	MA	92.19
9	02	83.26	2003-08-25	Pasadena	CA	90.13
10	04	98.57	2003-12-01	San Francisco	CA	106.95

16. [Update Raw XML in Alteryx Macros](#)

17. [Getting Started with Batch Macros](#)

18. Tools with an expression editor can be updated with questions:

The screenshot shows the Alteryx Designer interface with a workflow titled 'Workflow - Configuration'. The main area contains the following text:

Tools that contain an Expression Editor can be updated with questions

All tools can be updated with actions

A red arrow points from the text 'Tools that contain an Expression Editor can be updated with questions' to a 'List Box' tool icon in the palette. The palette also includes icons for Action, Check Box, Condition, Control Parameter, Date, Drop Down, Error Message, File Browse, Folder Browse, Macro Input, Macro Output, Map, Numeric Up Down, Radio Button, and Text Box.

The workflow consists of several connected tools: a blue 'List Box' tool, followed by a blue 'Macro Input' tool with an 'A' symbol, a red 'Macro Output' tool with a sigma symbol (Σ), and a blue 'Text Box' tool with three dots.

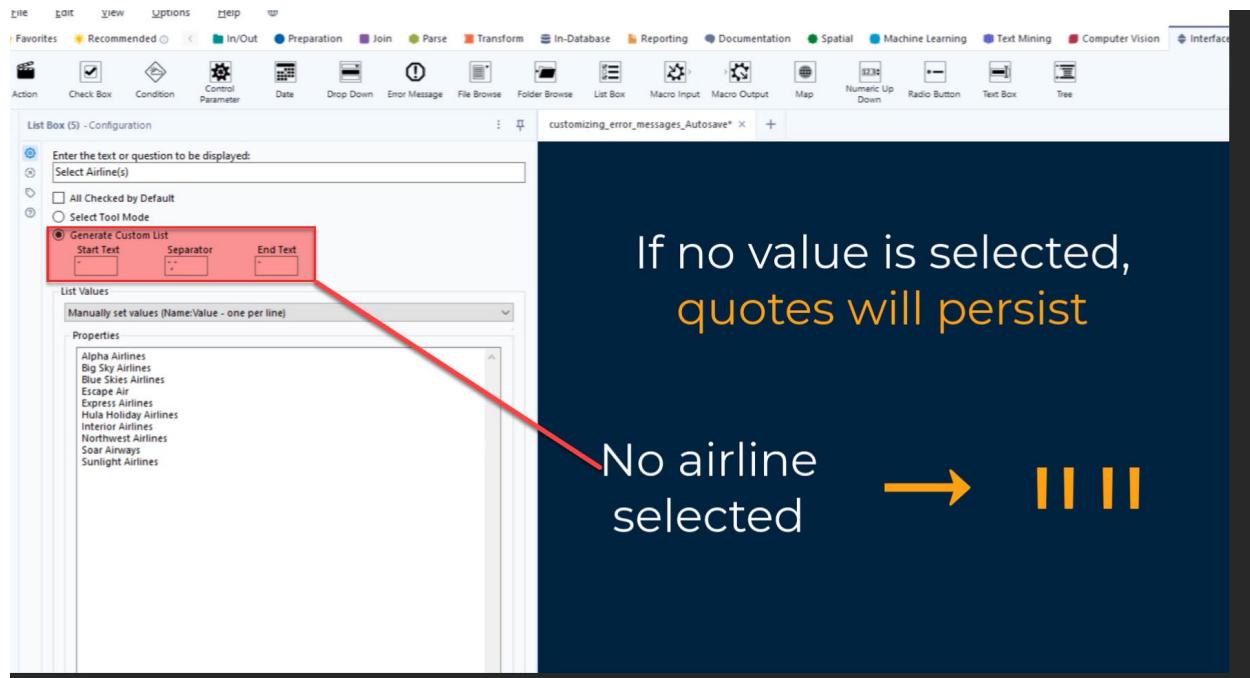
19. List Box questions:

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:

To generate, for example, a file with `Name` and `Value` fields, you can (in some cases) use a Summarize tool and group by twice to get the name and value. This is the case when the name and value are the same (e.g., displaying "Los Angeles" to the user and the value in the workflow is also "Los Angeles")

Name
Value in the second column
Then, click on the screen

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:

The screenshot shows the 'Trade Area (3) - Configuration' dialog. In the 'SpatialObject Field of Point Source' section, 'SpatialObj' is selected from a dropdown menu. A tooltip indicates: 'If a polyline or polygon is selected, its centroid will be used to create the trade area.' A red arrow points from this tooltip to the 'SpatialObj' dropdown. To the right, a process diagram shows a book icon connected to a target icon, with the text 'mechanic_locations.yxdb' below it.

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.

The screenshot shows the 'Trade Area (3) - Configuration' dialog. In the 'Radius, Doughnuts or Drivetime' section, 'Specific Value' is selected and '15,10,5' is entered into the input field. A tooltip states: 'The numbers must be entered in descending order, separated by commas, then select the screen.' A blue arrow points from this tooltip to the input field. To the right, a process diagram shows a book icon connected to a target icon, with the text 'mechanic_locations.yxdb' below it. Next to it is a box containing '15,10,5 Mi'.

Trade Area (3) - Configuration

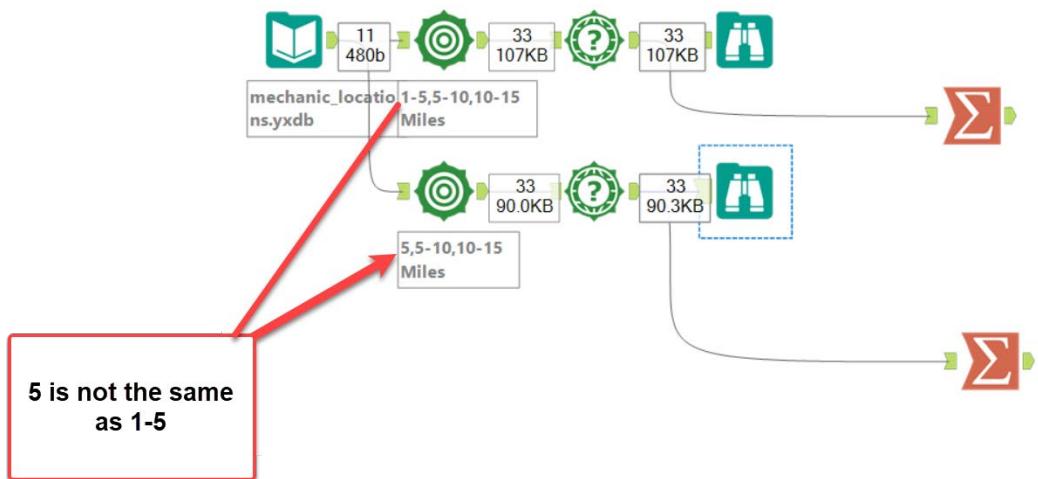
SpatialObject Field of Point Source
SpatialObj Include in Output

Radius, Doughnuts or Drivetime
 Specific Value:
Type: 5, 5-10, 10-15
 From Field:
Location

Units
 Radius (Miles)
 Radius (Kilometers)
 Drivetime Minutes
 Dataset:
No valid datasets found

Eliminate Overlap
(available for Specific Value Radius only)

Constructing Trade Areas.yxmd* +



4. [Dimension](#)
5. Spatial object details from the Browse tool

Results - Browse (2) - Input

2 of 2 Fields | Cell Viewer | 2 records displayed, 9,493 bytes | ↑ ↓

Record #1, Field Conservation Area SpatialObj (SpatialObj)

Type: Polygon
 Centroid: 34.3633918636904, -100.818636002479
 Number of Parts: 6
 Number of Points: 408
 Area: 21177.4243494988 Sq Mi / 54849.2772727421 Sq Km
 Length: 2662.32809472032 Mi / 4284.60174526958 Km
 Max Lat/Long: 36.4997, -99.133633
 Min Lat/Long: 31.693941, -103.042224

This is how you can view details about a spatial object

Record	Conservation Area Code	Conservation Area SpatialObj
1	26	Polygon - View Browse Tool Map Tab
2	29	Polygon - View Browse Tool Map Tab

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

Poly-Split (3) - Configuration

Spatial Field: Conservation Area SpatialObj

Split To:

- Points
- Regions All the spatial objects that are **not** holes
- Detailed Regions

Poly-Split (3) - Configuration

Spatial Field: Conservation Area SpatialObj

Split To:

- Points
- Regions
- Detailed Regions All the spatial objects, **including** holes

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
- 4
- 265
- 6

Results - Browse (2) - Input

2 of 2 Fields | Cell Viewer | 2 records displayed, 9,493 bytes

Record #1, Field Conservation Area SpatialObj (SpatialObj)

Type: Polygon
Centroid: 34.3633918636904, -100.818636002479
Number of Parts: 6
Number of Points: 408

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.

Question 3

If the following two spatial objects are combined, what will the output be?

- Null
- Line Spatial Object
- Point Spatial Object
- Polygon Spatial Object



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

Table Mode: Basic Pivot (CrossTab)

Station Name
 Station ID
 Longitude
 Latitude

Group By: **Grouping OFF**

Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Palmdale	Antelope Valley	1,829	2,100	\$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 Bernardino	2,392	3,233	\$28
Montalvo	Ventura County	1,169	1,330	\$15

Per Column Configuration

Station Name
 Station ID
 Longitude
 Latitude

Grouping ON

Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Palmdale	Antelope Valley	1,829	2,100	\$8
Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
West Corona	Inland Empire	788	1,015	\$13
Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Union Station	Los Angeles	13,645	16,097	\$14
Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Commerce	Orange County	1,982	2,096	\$17
Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Riverside-Downtown	Riverside	2,283	1,496	\$26
Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Montclair	San Bernardino	2,392	3,233	\$28
Station Name	Metro Line	Daily Avg Ridership	Previous Year Daily Avg Ridership	Average Ticket Purchase
Montalvo	Ventura County	1,169	1,330	\$15

2. Report Map

Report Map (6) - Configuration

Preview (Click to enlarge)

Sample Data

Values available are between -99 and 100. A negative value zooms in and a positive value zooms out

Settings Data Layers Legend

Map Size (W x H): 8 by 6 inches

Resolution: 1x (96 dpi)

Scale: Miles

Reference Base Map: Alteryx Light

Background Color: R=253, G=254, B=255

Map Drop Shadow: No

Expand Extent: % w/ a Minimum Width of Miles

The screenshot shows a GIS application interface. On the left, there's a legend editor for a layer named "Metro Stations". The legend tree includes "Style", "Label", "Theme" (which is selected), and "Base Layers - Points, Lines, Polygons". Under "Theme", a dropdown menu is open with options: "Smart Tile", "Equal Records", "Equal Ranges", "Manual Tile", and "Unique Value". A red box labeled "Numeric fields" encloses the "Equal Records", "Equal Ranges", and "Unique Value" options. A red arrow points from this box to another red box labeled "Categorical fields" which encloses the "Smart Tile" and "Manual Tile" options. On the right, there's a "Results - Report Map (8) - Input - Metro Stations" viewer showing a table with one record: "Record": 1, "Station Name": Lancaster, "Station ID": LAN, "Longitude": -118.13628295682151.

3. Layout

The screenshot shows a "Layout (31) - Configuration" dialog on the left and its preview on the right. The configuration dialog has sections for "Layout Mode" (set to "Each Individual Record"), "Group By", "Include Source Fields in Output", "Layout Configuration" (Orientation: Horizontal, Layout Width: Percentage 100%, Layout Height: Automatic), and "Per Column Configuration" (checkboxes for Chart, Map, Table, Text, with "Table" checked). The preview on the right shows a table with four columns: RECORD, GROUP, DATE, and TABLE. The rows are colored: Row 1 (RECORD 1, GROUP A, DATE 03/24) has a brown background; Row 2 (RECORD 2, GROUP A, DATE 04/16) has a teal background; and Row 3 (RECORD 3, GROUP B, DATE 06/01) has an orange background. The word "table" is repeated in the TABLE column under each row.

RECORD	GROUP	DATE	TABLE
1	A	03/24	table
2	A	04/16	table
3	B	06/01	table

Layout (31) - Configuration

Layout Mode: Each Individual Record

Group By:

Include Source Fields in Output

Layout Configuration

- Orientation: Horizontal
- Layout Width: Percentage 100 %
- Layout Height: Automatic
- Border
- Separator
- Cell Padding: 0 pixels

Per Column Configuration

- Chart
- Map
- Table
- Text

Width: Automatic

Alignment (V): Middle

Alignment (H): Center

Fill Color

LAYOUT MODE: Each Group of Records

RECORD	GROUP	DATE	TABLE
1	A	03/24	table
2	A	04/16	table
3	B	06/01	table

Layout (31) - Configuration

Layout Mode: Each Individual Record

Group By:

Include Source Fields in Output

Layout Configuration

- Orientation: Horizontal
- Layout Width: Percentage 100 %
- Layout Height: Automatic
- Border
- Separator
- Cell Padding: 0 pixels

Per Column Configuration

- Chart
- Map
- Table
- Text

Width: Automatic

Alignment (V): Middle

Alignment (H): Center

Fill Color

LAYOUT MODE: All Records Combined

RECORD	GROUP	DATE	TABLE
1	A	03/24	table
2	A	04/16	table

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

In-Database and General

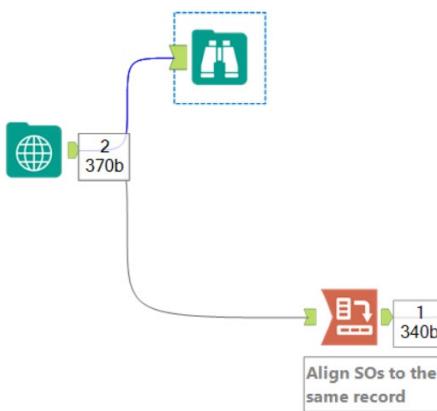
1. [How do Alteryx's In-Database tools work?](#)
2. [In-Database Workflows](#)
3. New line (i.e., \n) delimited data:

The screenshot shows a workflow titled "New Workflow1" in Alteryx Designer. The workflow consists of three inputs: "SurveyData.yxdb", "ContactData.yxdb", and "PhoneNumber.txt". The "SurveyData.yxdb" and "ContactData.yxdb" inputs are connected to a "Text to Columns" tool, which then connects to a "Text to Rows" tool. The "PhoneNumber.txt" input connects directly to a "Text to Rows" tool. The output from both rows tools connects to a "Union" tool, which finally connects to a "Table" tool. A red callout box highlights the "Text to Rows" tool from the "PhoneNumber.txt" input, containing the following text:
With data like this, it may be best to start by using the text to columns tool and split to rows on the "\n" delimiter. This is something you would typically do in Python with text files.

Results - Browse (S) - Input
Record #1, Field DownloadData (V.String)
1=V1,V2,V3,V4,V5,V6,V7,V8,V9,V10,V11,V12,LocationLatitude,LocationLongitude,LocationAccuracy,
ResponsesID,ResponseSetID,Name,ExternalDataDefinition,EmailAddress,IPAddress,Status,StartDate,EndDate,Finished,How fast is Alteryx?,Would you recommend Alteryx?,locationLatitude,locationLongitude
1_0tHFFGkzb1F5d1p,Default Response Set,Anonymous,,,2,2012-10-01 10:19:17,2012-10-01 10:19:18,1,5,5,,,-1,
2_3qL65QunPwCCEP,Default Response Set,Anonymous,,,2,2012-10-01 10:19:18,2012-10-01 10:19:18,1,2,4,,,-1,
3_40751514yjyP,Default Response Set,Anonymous,,,2,2012-10-01 10:19:18,2012-10-01 10:19:18,1,5,3,,,-1,
4_4D7y7s-nCT1Lz,Default Response Set,Anonymous,,,2,2012-10-01 10:19:20,2012-10-01 10:19:20,1,4,4,,,-1,
5_4hXfOg001H90CBL,Default Response Set,Anonymous,,,2,2012-10-01 10:19:21,2012-10-01 10:19:21,1,4,4,,,-1,
2_3ToYQkX3mpE5w1,Default Response Set,Anonymous,,,2,2012-10-01 10:19:21,2012-10-01 10:19:21,1,2,4,,,-1,

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

Before:



Results - Browse (46) - Input

Record	Label	SpatialObj
1	End	Polygon - View Browse Tool Map Tab
2	Start	Polygon - View Browse Tool Map Tab

After:

Cross Tab (64) - Configuration

Select data to transform.

Group data by these values:

- Label
- SpatialObj

Select All

Change Column Headers

Label

Values for New Columns

SpatialObj

Method for Aggregating Values

- First
- Last

Select All

New Workflow1* challenge_228_start_file.yxmd* × SelectRecords.yxmc* × + ...

position relative to the river and combine SO

Lookup Table

Results - Cross Tab (64) - Output

Record	End	Start
1	Polygon - View Browse Tool Map Tab	Polygon - View Browse Tool Map Tab

5. Operator syntax

Operator Syntax

AND = &&

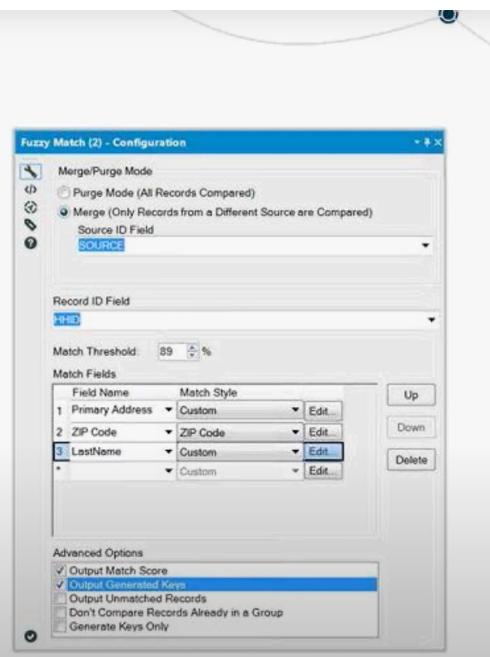
OR = ||

NOT = !

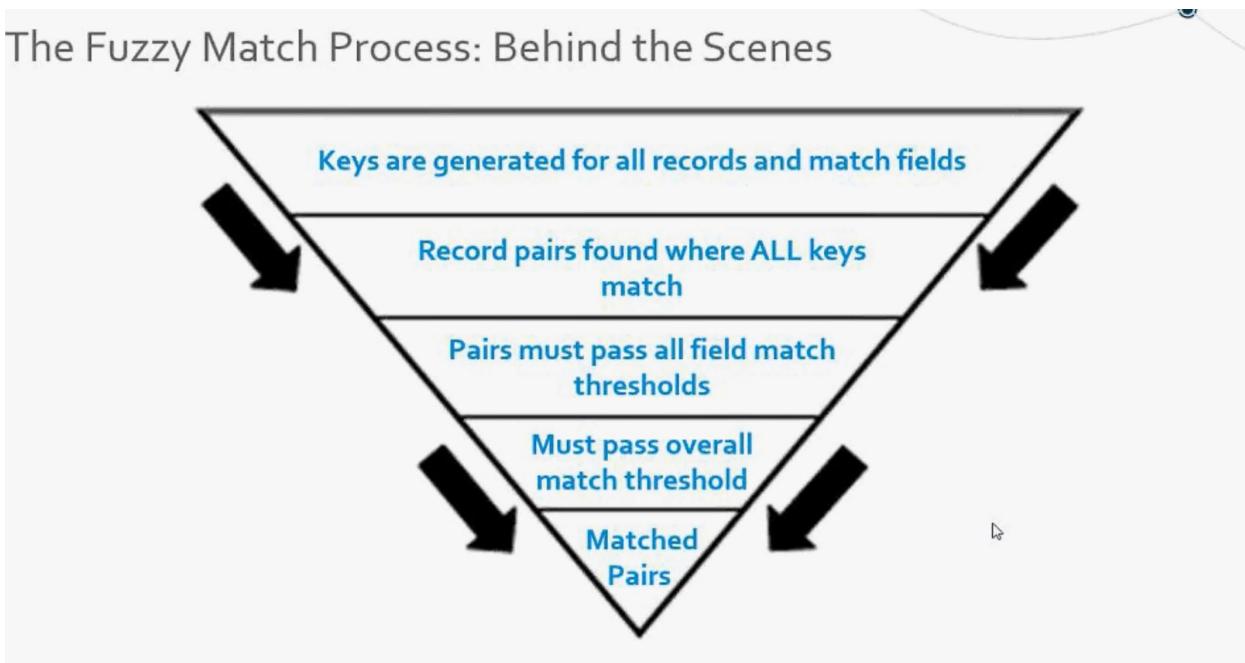
6. [Dynamic Input training](#)
7. [Running Total and Weighted Average](#)
8. [Parsing XML](#)
9. Fuzzy Matching

Fuzzy Matching Overview

- Deduping? Then Purge. Joining? Then Merge.
- Data preparation is the key to success
- The entire process is a series of ordered tests
 1. Key generation and match
 2. Match function field thresholds
 3. Overall match threshold
- Only the RecordID for matching records is output by default; MatchScore and MatchKey optional
- Always start with pre-configured Match Styles



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	TFISNKL
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:

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