

Is It Corked? Wine Machine Learning Predictions with OAC

Francesco Tisiot - @Ftisiot
Analytics Tech Lead - Rittman Mead

rittmanmead 
A DATA AND ANALYTICS COMPANY

Agenda

- OAC
- Data Scientist
- Become a Data Scientist



Oracle Analytics Cloud

- Platform Services (PaaS)
- Delivered entirely in the cloud:
 - No infrastructure footprint
 - Flexibility
 - Simplified, metered licensing
- Several options to suit your needs:
 - BYOL
 - Functionality bundled into 2 editions
 - Professional
 - Enterprise



Functions

OAC supports **Every** type of analytics

Classic



Modern



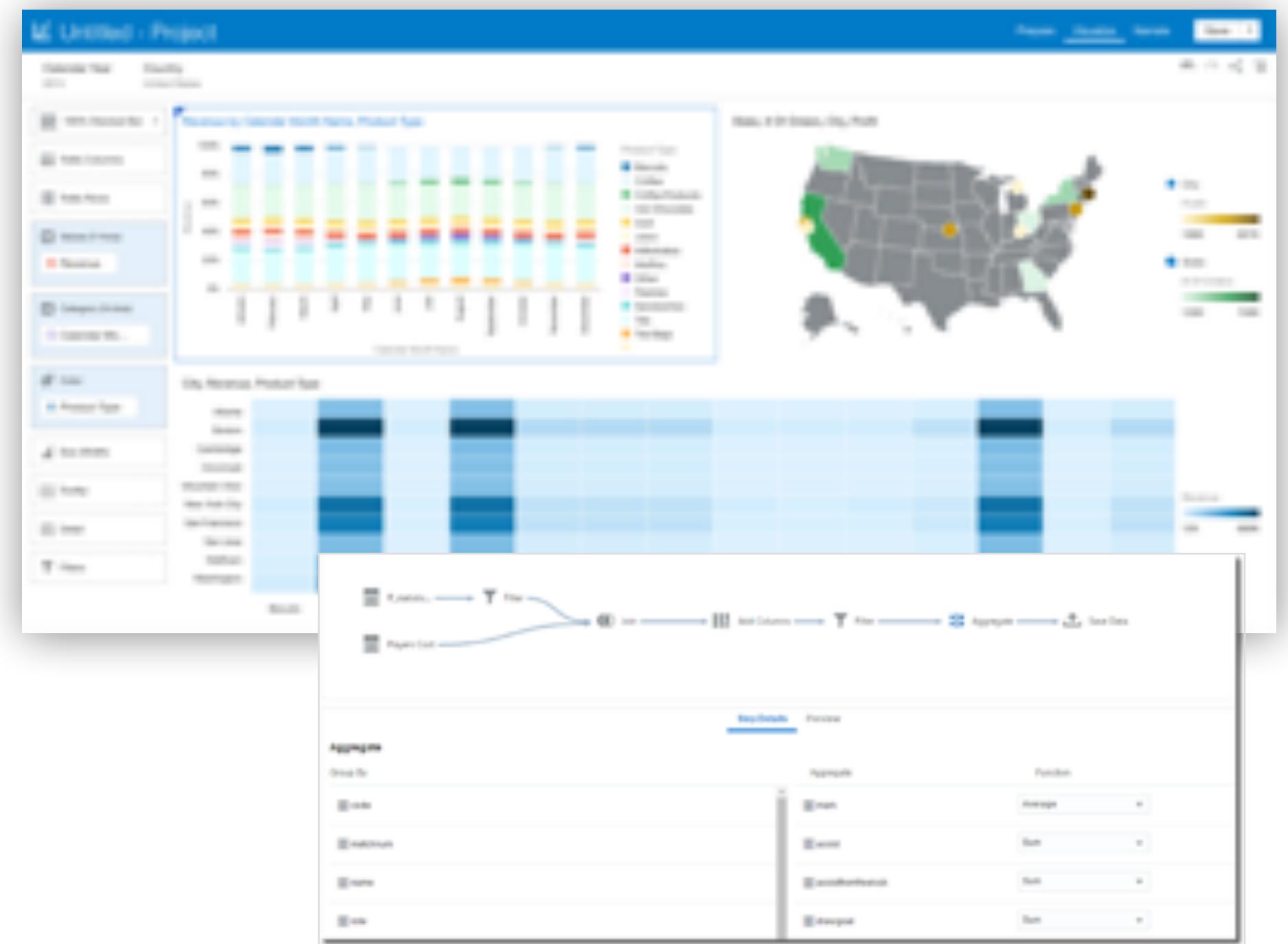
Classic Enterprise BI

- Similar to OBIEE 12c
 - Centrally maintained & governed
 - Semantic model
- Interactive Dashboards
 - KPI measurement & monitoring
 - Guided navigation paths
- BI Publisher
 - Highly formatted, burst outputs
- Action Framework
 - Navigation actions
 - Scheduled agents



Modern Data Discovery

- Data Preparation
 - Acquire data
 - Clean/Enrich
 - Transform
 - Repeatable Flows
- Data Visualisation
 - Create visual insights rapidly
 - Construct narrated storyboards
 - Share findings



Unified Analytics

Centralised
Reporting

Free
Discovery

Specific
Access
Control

Unique
Source of
Truth

Raw Data
To Insights

Data
Enrichment
and
Cleaning

Augmented Analytics

Data Enrichment

Suggestions

Natural
Language
Processing

Explain

One-Click
Advanced
Analytics

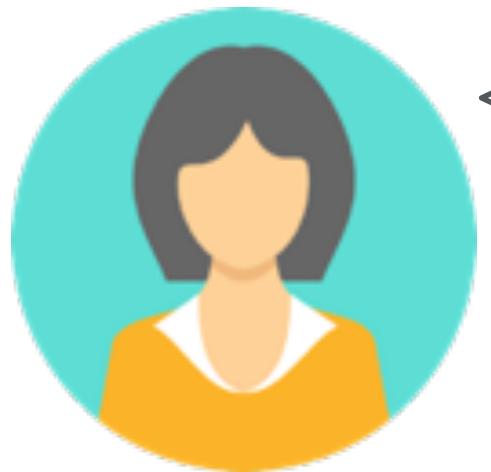
Advanced
Machine
Learning



OAC and Data Science



Basic Operations



What are the
Drivers for My
Sales?



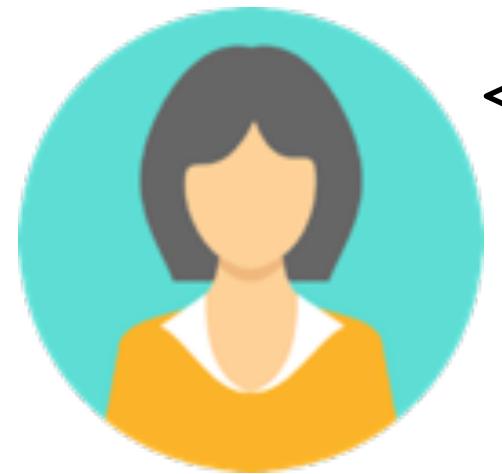
Based on my Experience
I can Guess....



Statistically Significant
Drivers for Sales Are ...

Augmented
Analytics

Basic Operations



YES/NO

50%

Basic ML
Model

70%

Before Starting.... Define the Problem!





Problem Definition: Predicting Wine Quality

Rule Based

Italy or France -> Good

Rest of the World -> Bad

Price \geq 10 Euros -> Good

Price $<$ 10 Euros -> Bad

Price $>$ 30 & Production Zone = Veneto & -> 6.5

TEP

Task

Estimate Wine
Good/Bad

Experience

Corpus of Wines
Descriptions with Ratings

Performance

Accuracy



Accuracy

		Predicted Value	
		Good	Bad
Real Value	Good		
	Bad		

Accuracy =  / ( + )

Dataset

The screenshot shows a dataset page on Kaggle. The title is "Wine Reviews" with a subtitle "139k wine reviews with variety, location, winery, price, and description". It includes a photo of purple grapes and a bio for user jackmills. Below the header are tabs for Data, Overview, Kernels (0), Discussions (0), and Activity. A "Download (57 MB)" button is prominent. The main content area has sections for Data (139k rows), Data Sources, About this file, and Columns.

Data (139k rows)

Data Sources

- (1) winemag-data-139k.csv (139k x 13)
- (1) winemag-data-failed.csv (139k x 11)
- (1) winemag-data-failed-wd.json

About this file

Here is a CSV version of the data I scraped. This dataset has three new fields -- Date (which you can parse the vintage from), Taster Name, and Taster Twitter Handle. This should also fix the duplicate entries problem in the first version of the dataset and add 129k unique reviews to play with.

Columns

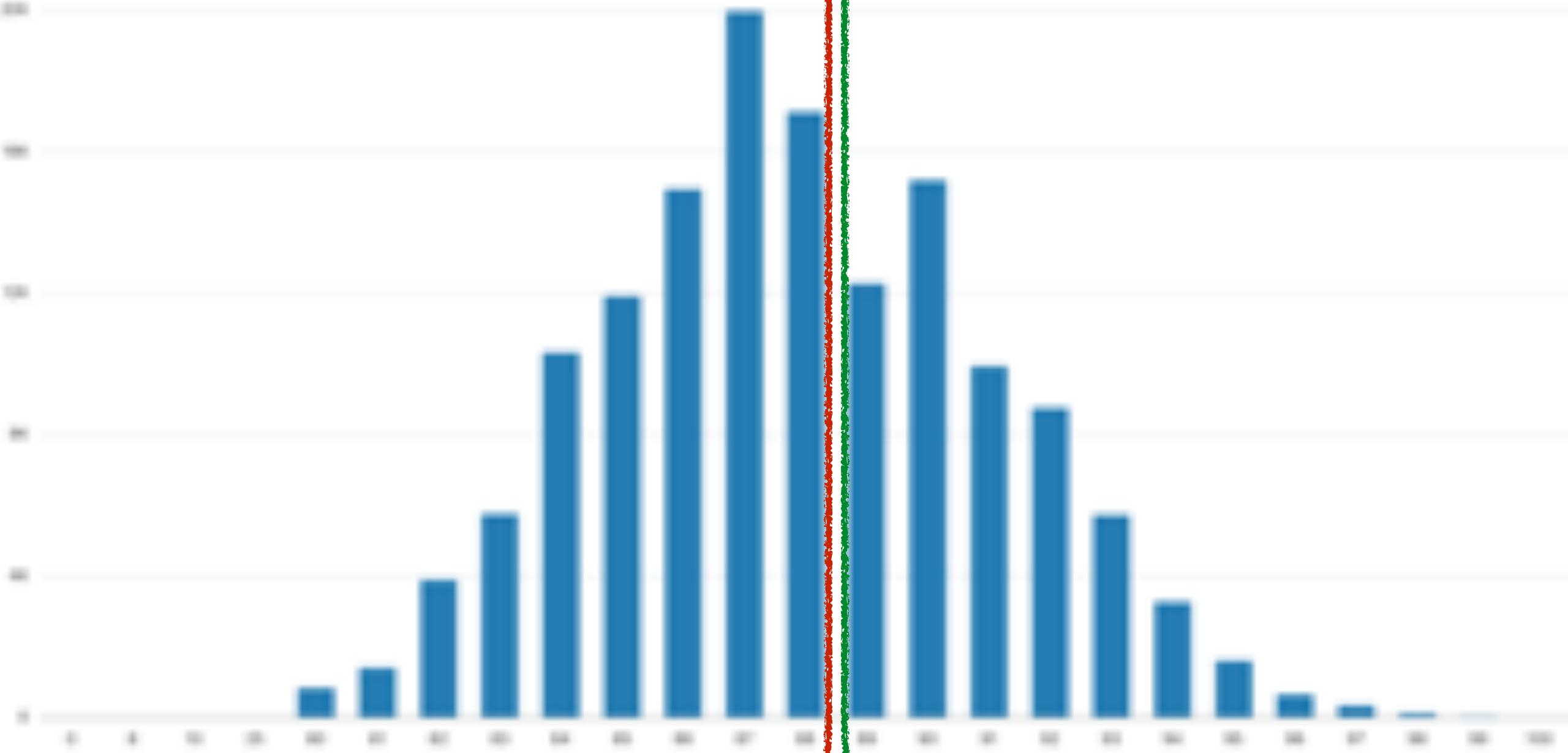
- (1) **points** The number of points WineEnthusiast rated the wine on a scale of 1-100 though they say they
- (1) **country** The country that the wine is from
- (1) **description**
- (1) **designation** The vineyard within the variety where the grapes that make the wine are from

The Data

A	B	C	D	E	F	G	H	I	J	K	L
#	country	description	designation	points	winery	region_1	region_2	region_3	variety	winery	
1	US	This tremendous 2009 Martha's Vineyard		98		2009 California	Napa Valley	Napa	Cabernet Sauvignon	Mata	
2	Spain	Ripe aromas of fig, iso-Caryolanum Sefuencio		98		2009 Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodriguez	
3	US	Mac Mawson's Honey & Special Selected Late		98		2009 California	Kingsville Valley	Sonoma	Sauvignon Blanc	MacMau	
4	Spain	This spent 20 months Reserve		98		09 Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ajani	
5	France	This is the top wine from La Bruffe		98		09 Provence	Rasteau		Roussanne red blend	Domaine de la Bruffe	
6	Spain	Deep, dense and pure Rumania		98		79 Northern Spain	Toro		Tinta de Toro	Rumania	
7	Spain	Slightly gritty taste from Ben Romo's		98		69 Northern Spain	Toro		Tinta de Toro	Macrodia	
8	Spain	Lush, earthy black-fruit (Cardeñaz 1996)		98		09 Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodriguez	
9	US	This re-named Pinot Noir		98		69 Oregon	Chehalem Mountains	Willamette Valley	Pinot Noir	Bengtorells	
10	US	The producer sources Gasp's Etched Vineyard		98		69 California	Sonoma Coast	Sonoma	Pinot Noir	Blue Farm	
11	Italy	Elegance, complexity, Rosso delle Chiese		98		89 Northeastern Italy	Collio		Primitivo	Borgo del Tiglio	
12	US	From 28-year-old vine Ekuus Vineyard Wines		98		49 Oregon	Ribbon Ridge	Willamette Valley	Pinot Noir	Patricia Green Cellars	
13	US	A standout even in the winter vineyard		98		49 Oregon	Oundee Hills	Willamette Valley	Pinot Noir	Patricia Green Cellars	
14	France	This wine is in peak no Château Merlus Pre		98		99 Southwest France	Madiran		Tannat	Vignobles Brumont	
15	US	With its sophisticated Grace Vineyard		98		2009 Oregon	Oundee Hills	Willamette Valley	Pinot Noir	Domaine Serene	
16	US	First made in 2006, no Syrah		98		99 Oregon	Willamette Valley	Willamette Valley	Chardonnay	Bengtorells	
17	US	This blockbuster, pure Rame Vineyard		98		3009 California	Diamond Mountain-D-Ridge	Cabernet Sauvignon	Hall		
18	Spain	Nicely aged blackberry & citrus Reserve Pin		98		89 Northern Spain	Ribera del Duero		Tempranillo	Villanera	
19	France	Coming from a steep Le Pigeonnier		98		2009 Southwest France	Cahors		Malbec	Château Lagrézette	
20	US	This fresh and fruity in Gasp's Etched Vineyard		98		79 California	Sonoma Coast	Sonoma	Pinot Noir	Gary Farrell	

Bad

Good



Become a Data Scientist with OAC

Connect

Clean

Transform
&
Enrich

Analyse

Train
&
Evaluate

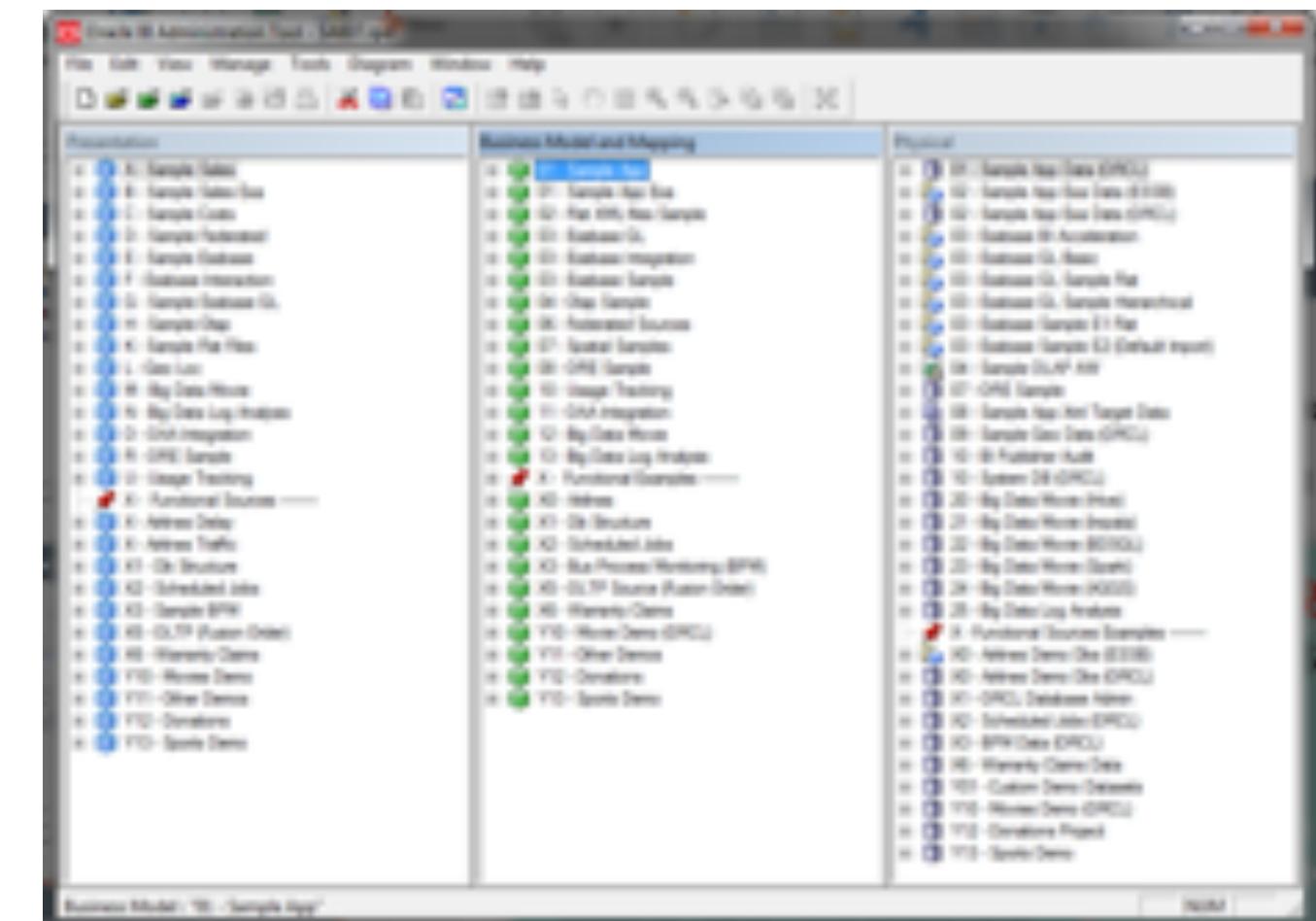
Predict

Connection Options in OAC

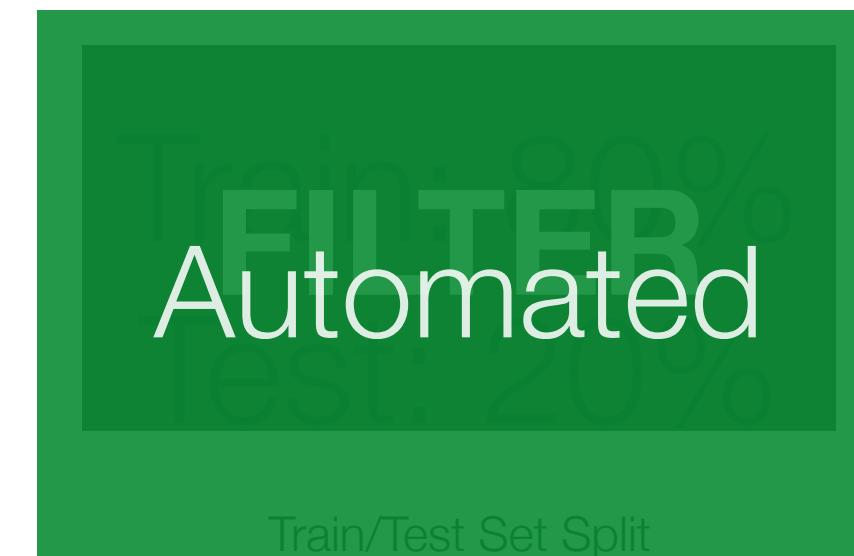
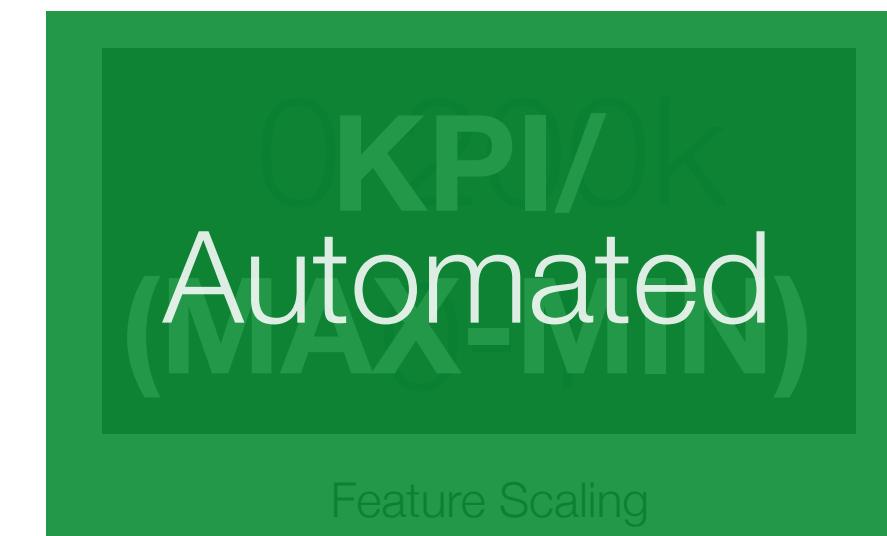
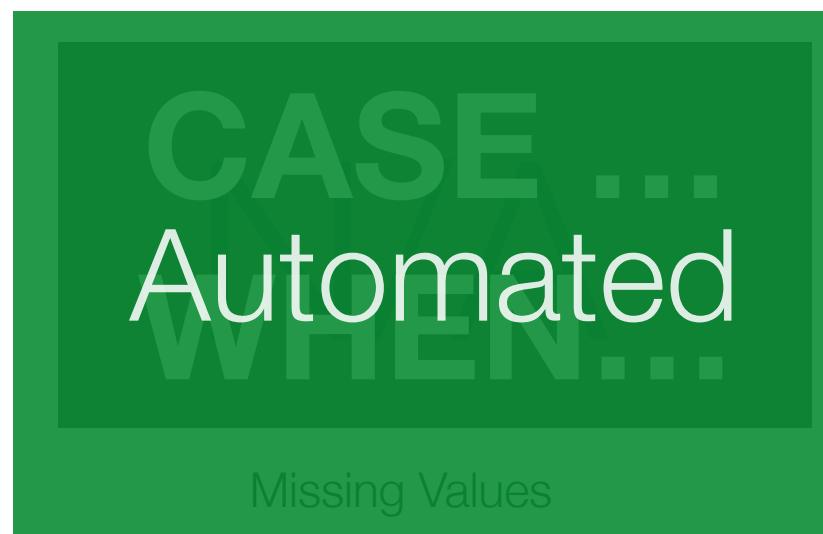


External
Data Sources

Pre-Defined
Data Models



Cleaning What?



Aggregation

Train/Test Set Split

Feature Engineering

Location -> ZIP Code

Additional
Data Sources?

Name -> Sex

2 Locations -> Distance

Data Flow

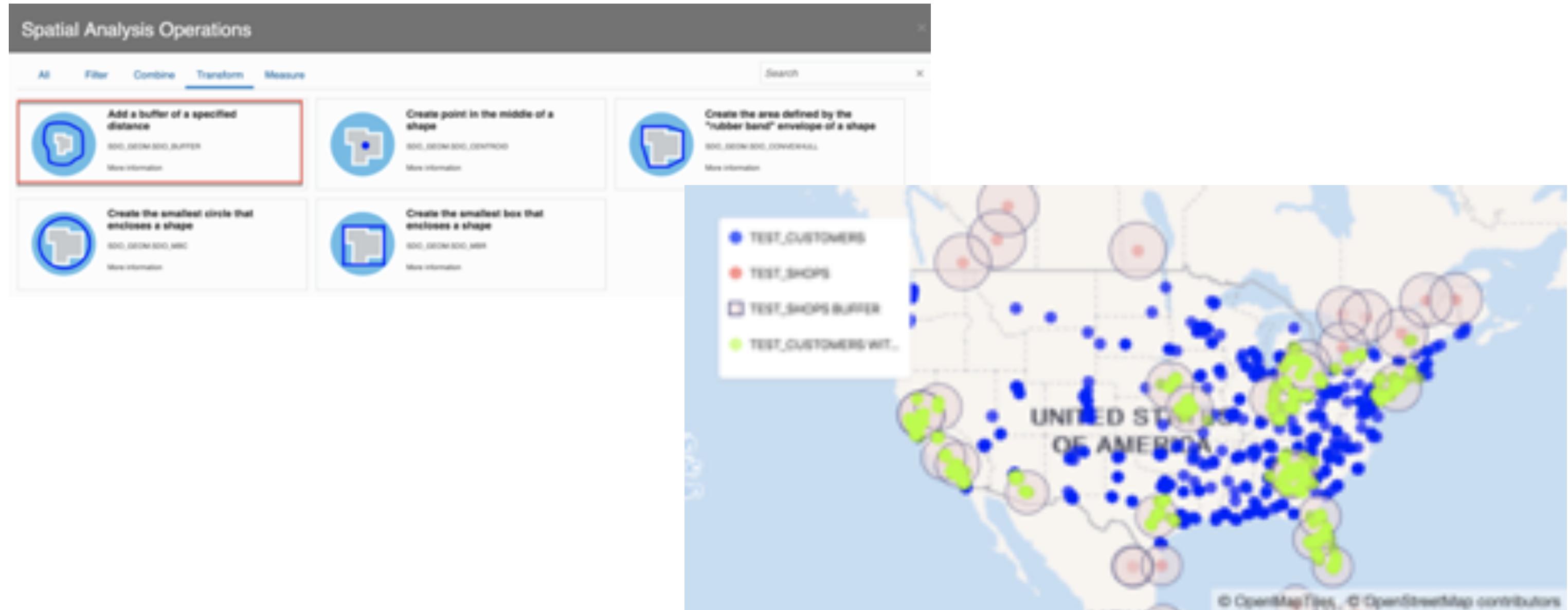
Day/Month/Year -> Date

Data Preparation Recommendations

The screenshot shows a data preparation interface with several panels:

- Left Sidebar:** Labeled "Preparation Steps" and "0 / 0". It contains a list of steps:
 - Import countries from JSON (Completed)
 - Import cities from JSON (Completed)
 - Import country with cities (Completed)
 - Import cities with countries (Completed)
 - Import cities (Completed)A "Ready script" button is at the bottom.
- Top Bar:** Includes "Exclusions", "New", "Formulas", "Data Project", and a dropdown menu "Healthy (19)".
- Tables:** Three tables are visible:
 - country:** Columns: ID, name. Rows include Germany, France, Austria, US, DE, Spain, France, US, CA, United States, US, United States, United States, Portugal, Greece.
 - country_country_name:** Columns: ID, name. Rows include Germany, France, Austria, United States, United States, Spain, France, United States, United States, United States, United States, Portugal, Greece.
 - country_id:** Columns: ID, name. Rows include DEU, FRA, AUT, USA, USA, USA, USA, USA, USA, USA, USA, GRC.
- Right Panel:** A dropdown menu titled "Healthy (19)" containing 19 items, each starting with "Import country with ...":
 - Import country with JSON
 - Import country with cities, names
 - Import country with cities
 - Import country with cities, names
 - Import country with names
 - Import country with regions, IDs
 - Import country with population
 - Import country with countries
 - Import country with IDs
 - Import country with currency, IDs
 - Import country with currency, names
 - Import country with density, names
 - Import country with density, IDs
 - Import country with names, IDs
 - Import country with regions, names
 - Import country with regions, IDs
 - Import country with regions, names, IDs
 - Import country with regions, names, IDs, names

Spatial Enrichment



Oracle Spatial Studio

<http://ritt.md/spatial-studio>

Data Overview

Results

Data Element	Data Type	Read As	Aggregation	Sample Value
#	number	▲ Attribute	none	100, 800, 1000, 8000, 10000, 100000, 800000
country	varchar(5)	▲ Attribute	none	US, France, Italy, Spain, Portugal, Germany, Argentina, Chile, Austria, Greece
country_continent	varchar(50)	▲ Attribute	none	North America
country_flag	varchar(50)	▲ Attribute	none	USA, UK, FR
country_id	varchar(50)	▲ Attribute	none	USA, FR, UK
country_idc_number	number	● Measure	sum	1000, 1000, 1000
country_idc	varchar(50)	▲ Attribute	none	US, UK, FR
description	varchar(50)	▲ Attribute	none	This elegant wine combines subtle nutmeg and cinnamon aromas with ripe apric...
designation	varchar(50)	▲ Attribute	none	Reserve, Estate Reserve, Reserve, Estate Bottled, Waller Riesling, Zweig, Zweig...
points	number	● Measure	sum	85, 85, 85, 87, 81, 85, 85, 85, 85, 85
price	varchar(5)	▲ Attribute	none	15, 15, 15, 15, 15, 15, 15, 15, 15
region	varchar(5)	▲ Attribute	none	California, Oregon, Sonoma, Napa Valley, Mendocino, Washington, Northern Idaho, M...
region_1	varchar(5)	▲ Attribute	none	Willamette Valley, Napa Valley, Sonoma, Mendocino, Willamette River Valley, Willam...
region_2	varchar(5)	▲ Attribute	none	Central Coast, Sonoma, Willamette Valley, Napa, Columbia Valley, Mendocino, L...
winery	varchar(5)	▲ Attribute	none	Hest Hest, (Hawthorn) Sonoma, says that Hest (Gaston Bourgeon), has the...
winery	varchar(5)	▲ Attribute	none	Saint-Honoré Hill, St. Honoré, Bergoth, Dr. Hendrik de Poorter, Rosack, Saint-Honoré...

Analyse - Explain

points

- Add to Selected Visualization
- Create Best Visualization
- Pick Visualization...

- Create Filter
- Explain points**

Explain points

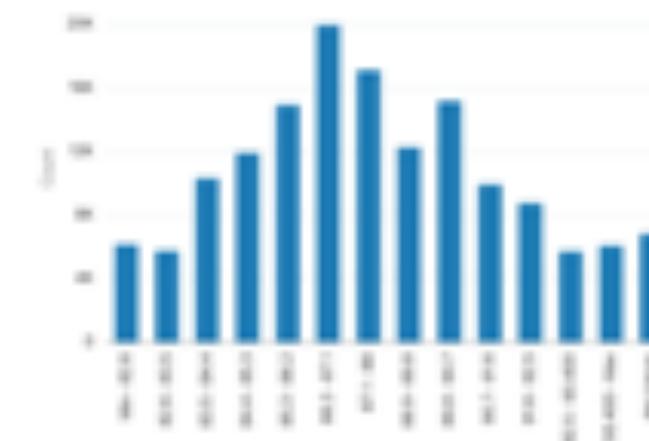
Basic Facts about points

What are the values of points and how do they relate to each other?

Key Drivers of points

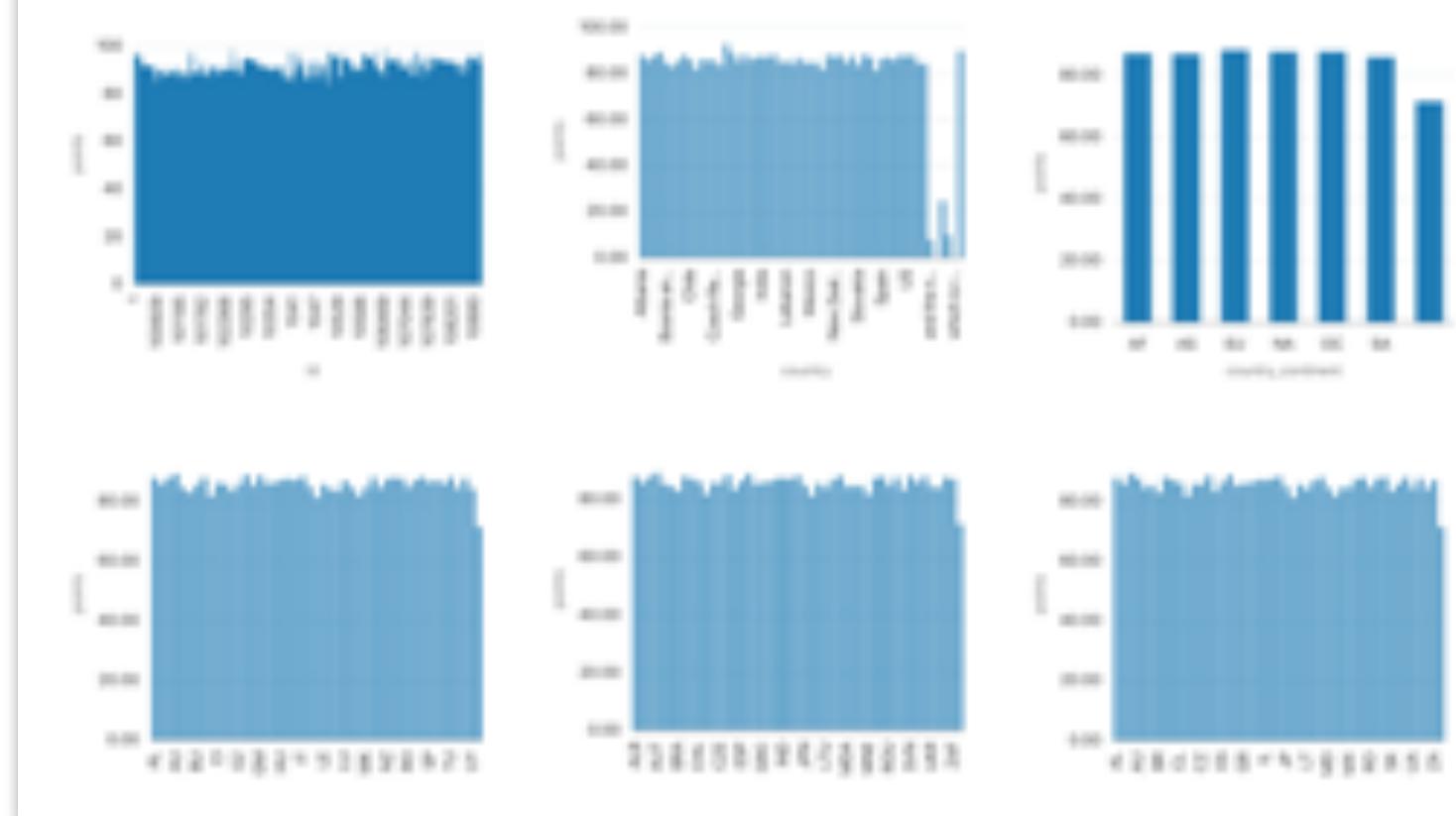
What elements in the data best explain the values of points?

Basic facts about points



points is a Numeric Measure, whose average across 190,000 rows is 87.00. The values of points on each row range from 0.00 to 100.00 and is 87.00 on average.

The charts below summarize the values of points by the measures in this data set. Click the checkmarks above any of the visuals to add them to your project when done.

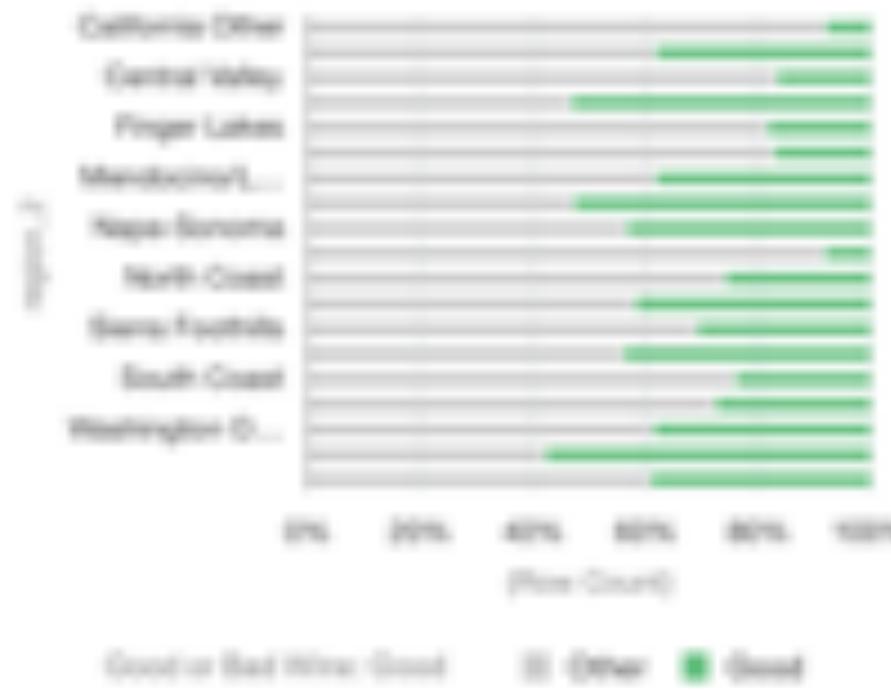
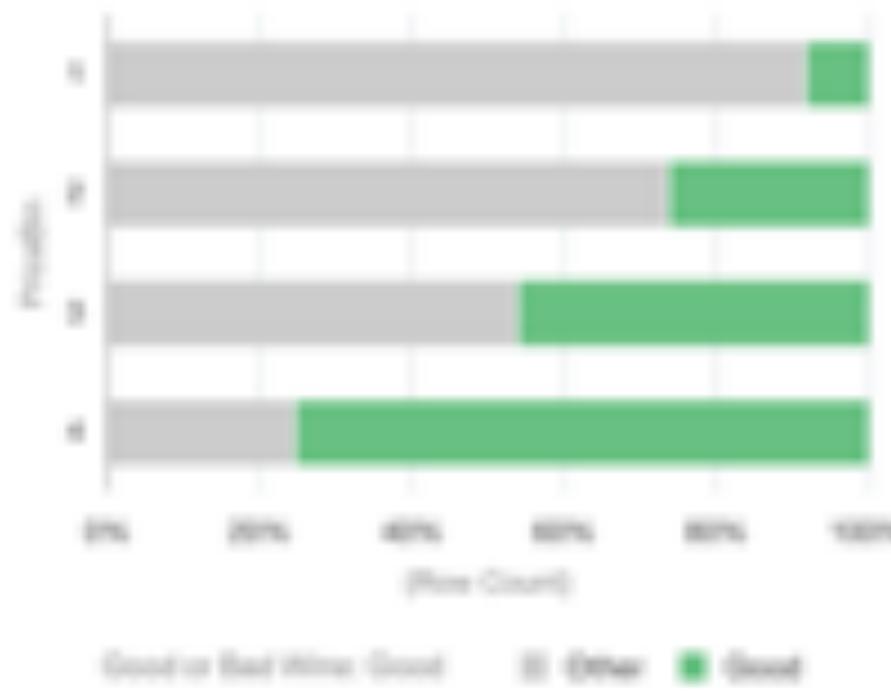


Explain - Key Drivers

Key Drivers of Good or Bad Wine

Based on Good or Bad Wine: **Good** the 2 attributes that are most strongly correlated are: Price/Unit, region_2:

The charts below show the distribution of Good or Bad Wine values across each of the key drivers. Click the checkboxes above any of the visuals to add them to your project when done.

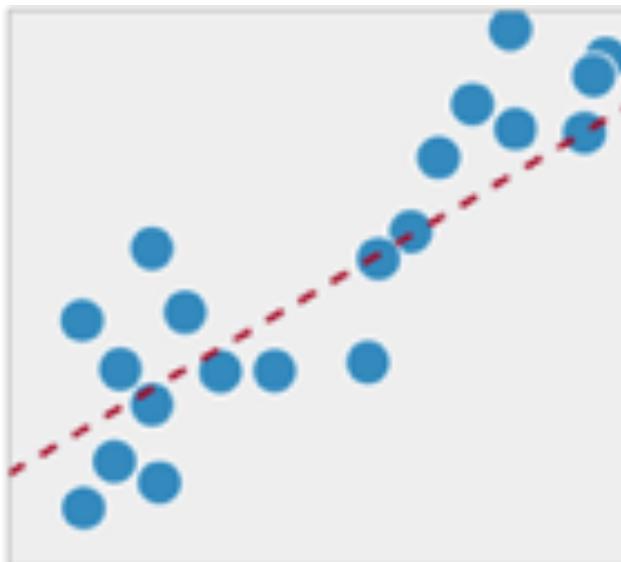


Train - What Problem are we Trying to Solve?

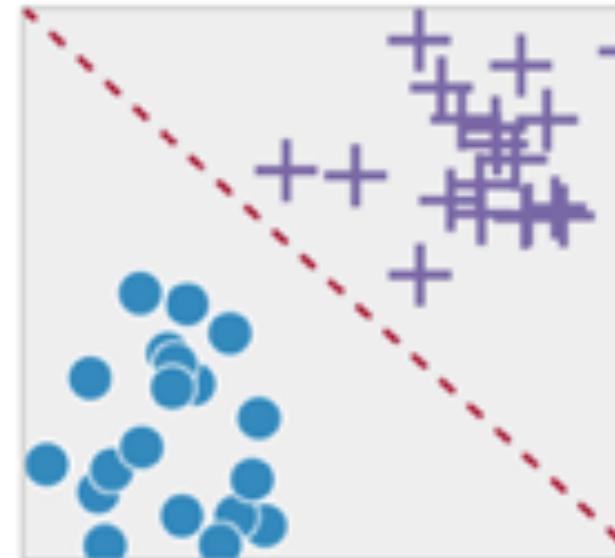
Supervised

“I want to predict the value of Y,
here are some examples”

Regression



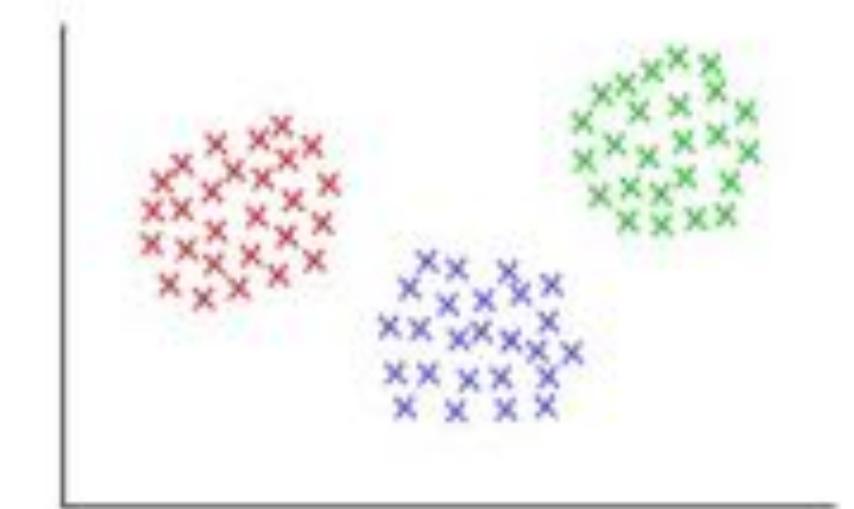
Classification



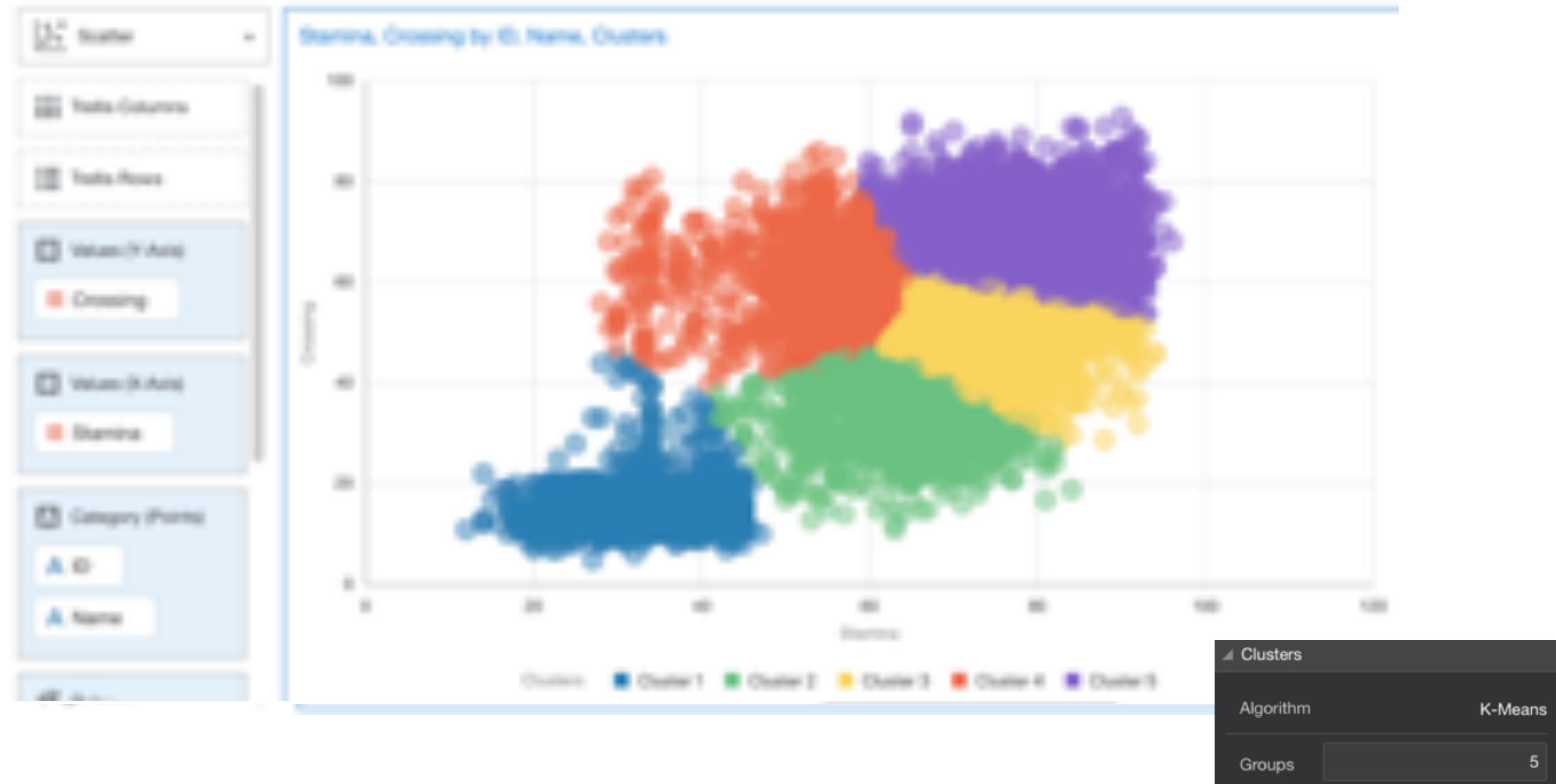
Unsupervised

“Here is a dataset,
make sense out of it!”

Clustering



Model Training - Easy Models



DataFlow Train Model



Select Train Numeric Prediction Model Script

Search

- Linear Regression for model training
- Elastic Net Linear Regression for model training
- Random Forest for Numeric model training
- CART for Numeric Prediction training

Which Model - Parameters To Pick?

Select Train Numeric Prediction Model Script

Search

- Linear Regression for model training
- Elastic Net Linear Regression for model training
- Random Forest for Numeric model training
- CART for Numeric Prediction training

Train Numeric Prediction

Model Training Script [Linear Regression for model training](#)

Target [Select a column](#)

Regression Method [Linear](#)

Regularization Weight [0](#)

Categorical Column Imputation [Most Frequent](#)

Numerical Column Imputation [Mean](#)

Categorical Encoding Method [Reference](#)

Minimum Red Value Percent [80](#)

Free Partition Percent [80](#)

Select, Try, Save, Change, Try, Save

The screenshot illustrates a machine learning workflow interface with the following components:

- Top Navigation:** A horizontal bar with four rounded rectangular buttons: "Working...", "Select Columns", "Train Numeric Prediction" (highlighted in blue), and "Save Model".
- Left Sidebar:** A "Train Numeric Prediction" section containing:
 - "Model Training Script": Set to "Elastic Net Linear Regression for model training".
 - "Target": "score" (input, the target variable for numeric prediction).
 - "L1 Ratio": "0.5" (radio button selected).
 - "L2 Ratio": "0.5" (radio button selected).
- Central Area:** A "Select Train Numeric Prediction Model Script" dialog box. It includes:
 - A search bar with a magnifying glass icon.
 - Four model icons: "Linear Regression for model training", "Elastic Net Linear Regression for model training", "Random Forest for Numeric model training", and "CART for Numeric Prediction training".
 - A navigation bar with tabs: "Data Sets", "Connections", "Data Flows" (selected), and "Sequences".
- Bottom Area:** A "Machine Learning" section with two tabs: "Scripts" (selected) and "Models". It lists three items:

Type	Name
»»	ELN1
»»	LR2
»»	LR1

Below this is another table:

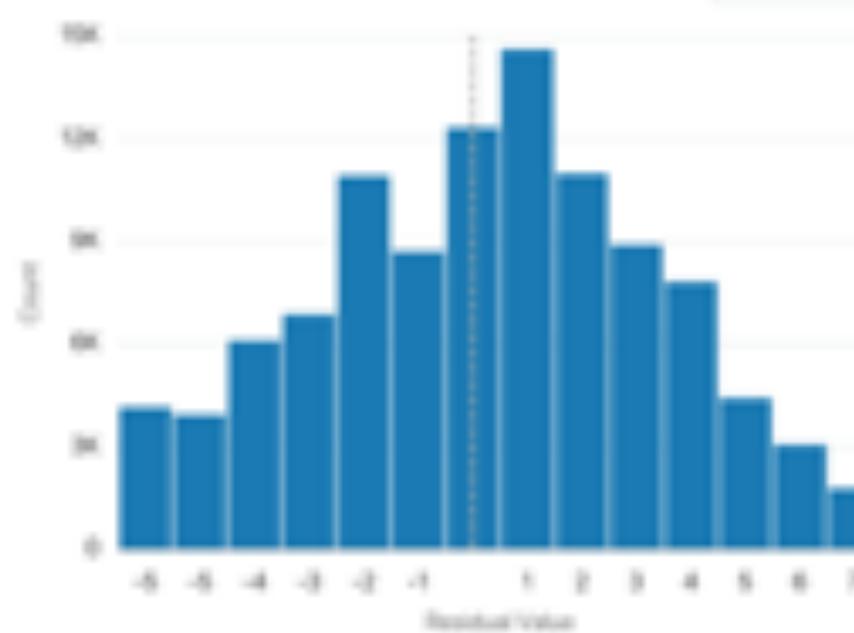
Type	Name
»	ELN1
»	LR2
»	LR1

Compare

LR1
Numeric Prediction Model

General Quality Related

Count by Residual Value Number of bins: 20



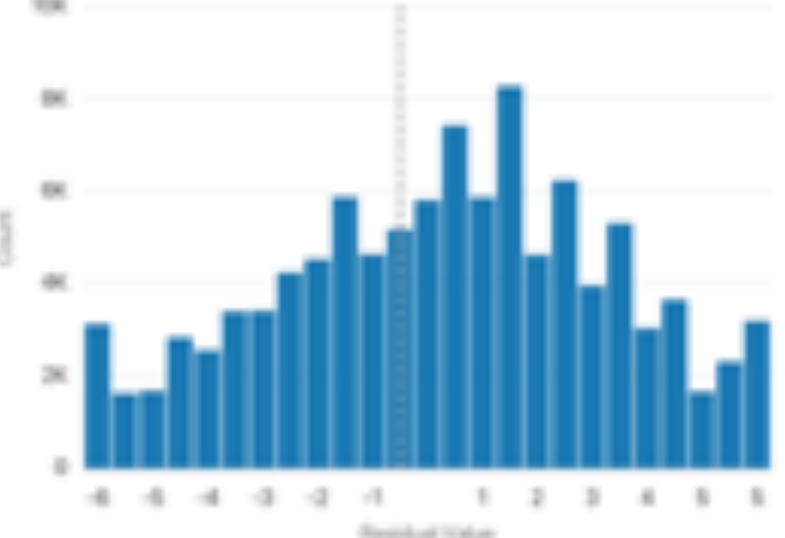
Mean Absolute Error (MAE): 3.66
Median Absolute Error: 3.34
Root Mean Squared Error: 3.23
Relative Absolute Error (RAE): 1.06
Relative Squared Error (RSE): 1.06
Coefficient of Determination (R^2): 49%

Close

LR2
Numeric Prediction Model

General Quality Related

Count by Residual Value Number of bins: 20



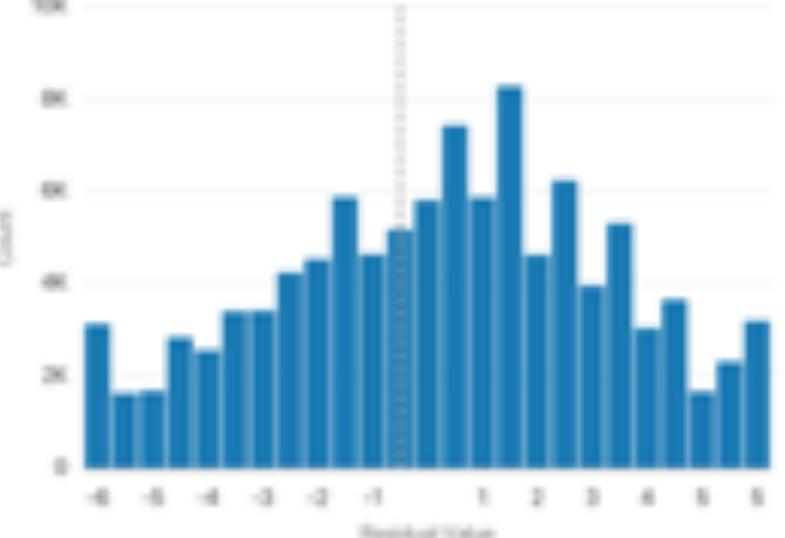
Mean Absolute Error (MAE): 3.56
Median Absolute Error: 3.20
Root Mean Squared Error: 3.16
Relative Absolute Error (RAE): 0.99
Relative Squared Error (RSE): 0.99
Coefficient of Determination (R^2): 49%

Close

LR3
Numeric Prediction Model

General Quality Related

Count by Residual Value Number of bins: 20



Mean Absolute Error (MAE): 3.56
Median Absolute Error: 3.20
Root Mean Squared Error: 3.16
Relative Absolute Error (RAE): 0.99
Relative Squared Error (RSE): 0.99
Coefficient of Determination (R^2): 49%

Close

Compare - Classification

		Predicted Values		
		0.0	1.0	Total
Actual Values	0.0	40439	471	40910 (90%)
	1.0	3761	866	4627 (10%)
	Total	44200 (97%)	1337 (3%)	45537 (100%)

Use On the Fly or with a Dataflow

Add Data Set...

Create Scenario...

Add Create Scenario - Select Model

Search

Type Name

- BinaryGant2
- BinaryGant1
- BinaryLogistic1
- BNF1
- LRF1
- LRF1

Edit Scenario - Map Your Data

Select which Data Set you want to use with the Model

Data Set PredictEvents

For each model input listed on the left, select a corresponding data element from your project

Model Input	Map To	Available Data Elements
bodypart	bodypart	Bodypart BodyPart BodyPart BodyPart BodyPart
location	location	Location Location Location Location Location
player	player	Player Player Player Player Player
situation	situation	Situation Situation Situation Situation Situation
is_goal	is_goal	IsGoal IsGoal IsGoal IsGoal IsGoal
+ Requested Fields		



Demo

Congratulations!



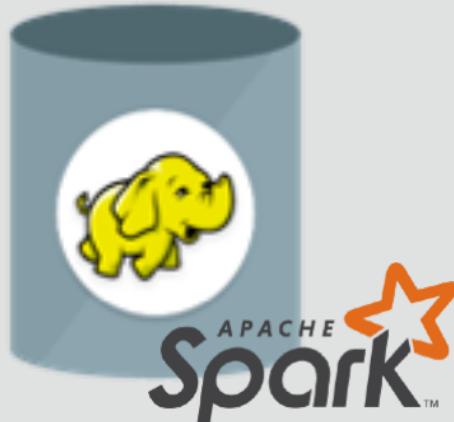
...You are now a Data Scientist!

ML Production Deployment

Data Scientist

ML -> Data

Oracle Machine Learning



Oracle Machine Learning

OML4SQL

Oracle Advanced Analytics
SQL API

OML4R

Oracle R Enterprise
R API

OML4Py*

Python API

OML Microservices*

Supporting Oracle Applications
Image, Text, Scoring, Deployment,
Model Management

OML Notebooks

with Apache Zeppelin on
Autonomous Database

Oracle Data Miner

Oracle SQL Developer extension

OML4Spark

Oracle R Advanced Analytics
for Hadoop

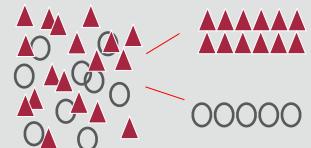


Oracle Machine Learning Algorithms



CLASSIFICATION

Naïve Bayes
Logistic Regression (GLM)
Decision Tree
Random Forest
Neural Network
Support Vector Machine
Explicit Semantic Analysis



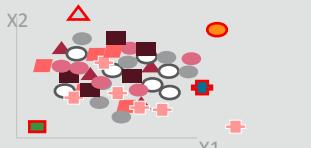
CLUSTERING

Hierarchical K-Means
Hierarchical O-Cluster
Expectation Maximization (EM)



ANOMALY DETECTION

One-Class SVM



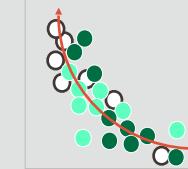
TIME SERIES

Forecasting - Exponential Smoothing
Includes popular models
e.g. Holt-Winters with trends,
seasonality, irregularity, missing data



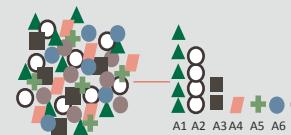
REGRESSION

Linear Model
Generalized Linear Model
Support Vector Machine (SVM)
Stepwise Linear regression
Neural Network
LASSO



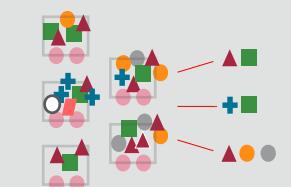
ATTRIBUTE IMPORTANCE

Minimum Description Length
Principal Comp Analysis (PCA)
Unsupervised Pair-wise KL Div
CUR decomposition for row & AI



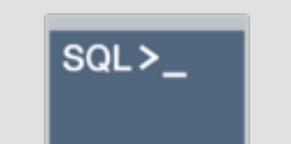
ASSOCIATION RULES

A priori/ market basket



PREDICTIVE QUERIES

Predict, cluster, detect, features

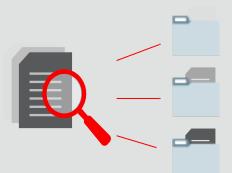


SQL ANALYTICS

SQL Windows
SQL Patterns
SQL Aggregates

FEATURE EXTRACTION

Principal Comp Analysis (PCA)
Non-negative Matrix Factorization
Singular Value Decomposition (SVD)
Explicit Semantic Analysis (ESA)



TEXT MINING SUPPORT

Algorithms support text
Tokenization and theme extraction
Explicit Semantic Analysis (ESA) for
document similarity



STATISTICAL FUNCTIONS

Basic statistics: min, max,
median, stdev, t-test, F-test, Pearson's,
Chi-Sq, ANOVA, etc.



R PACKAGES

Third-party R Packages
through Embedded Execution
Spark MLlib algorithm integration

MODEL DEPLOYMENT

SQL – 1st Class Objects
Oracle RESTful API (ORDS)
OML Microservices (for Apps)



Oracle Machine Learning

Machine Learning Notebook for Autonomous Data Warehouse Cloud

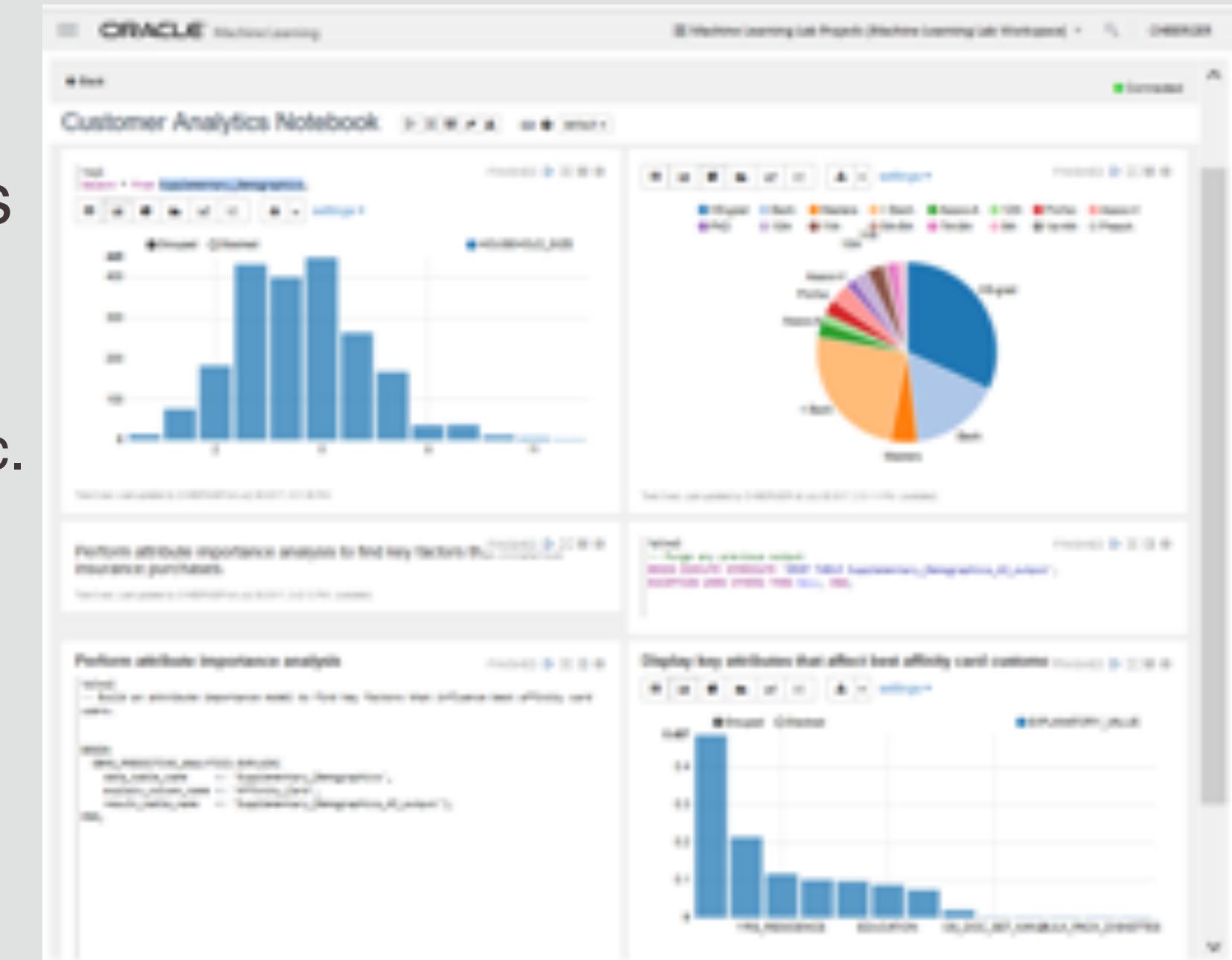


Key Features:

Collaborative UI for data scientists

Packaged with Autonomous Data Warehouse Cloud
Easy access to shared notebooks,
templates, permissions, scheduler, etc.
SQL ML algorithms API

Supports deployment of ML analytics



Become a Data Scientist with OAC



<http://ritt.md/OAC-datascience>

ML in Action with OAC



A dark green rectangular overlay containing white text. The text reads "Hello Wine, Goodbye Problems" in a sans-serif font, with each word on a new line.

<http://ritt.md/OAC-ML-Video>

Insights Lab

<https://www.rittmanmead.com/insight-lab/>