ODTUG Kscope19

SEATTLE, WASHINGTON • JUNE 23-27

PLEASE FILL OUT YOUR EVALUATIONS

SEATTLE

Washington State Convention Center

Is It Corked? Wine Machine Learning Predictions with OAC

Francesco Tisiot
BI Tech Lead at Rittman Mead

Francesco Tisiot

BI Tech Lead at Rittman Mead

Verona, Italy

Rittman Mead Blog

10 Years Experience in BI/Analytics

francesco.tisiot@rittmanmead.com

@FTisiot

Oracle ACE

About Rittman Mead

Rittman Mead is a **data and analytics company** who specialise in data visualisation, predictive analytics, enterprise reporting and data engineering.

We use our skill, experience and know-how to work with organisations across the world to interpret their data. We enable the business, the consumers, the data providers and IT to work towards a common goal, **delivering innovative and cost-effective solutions** based on our core values of thought leadership, hard work and honesty.

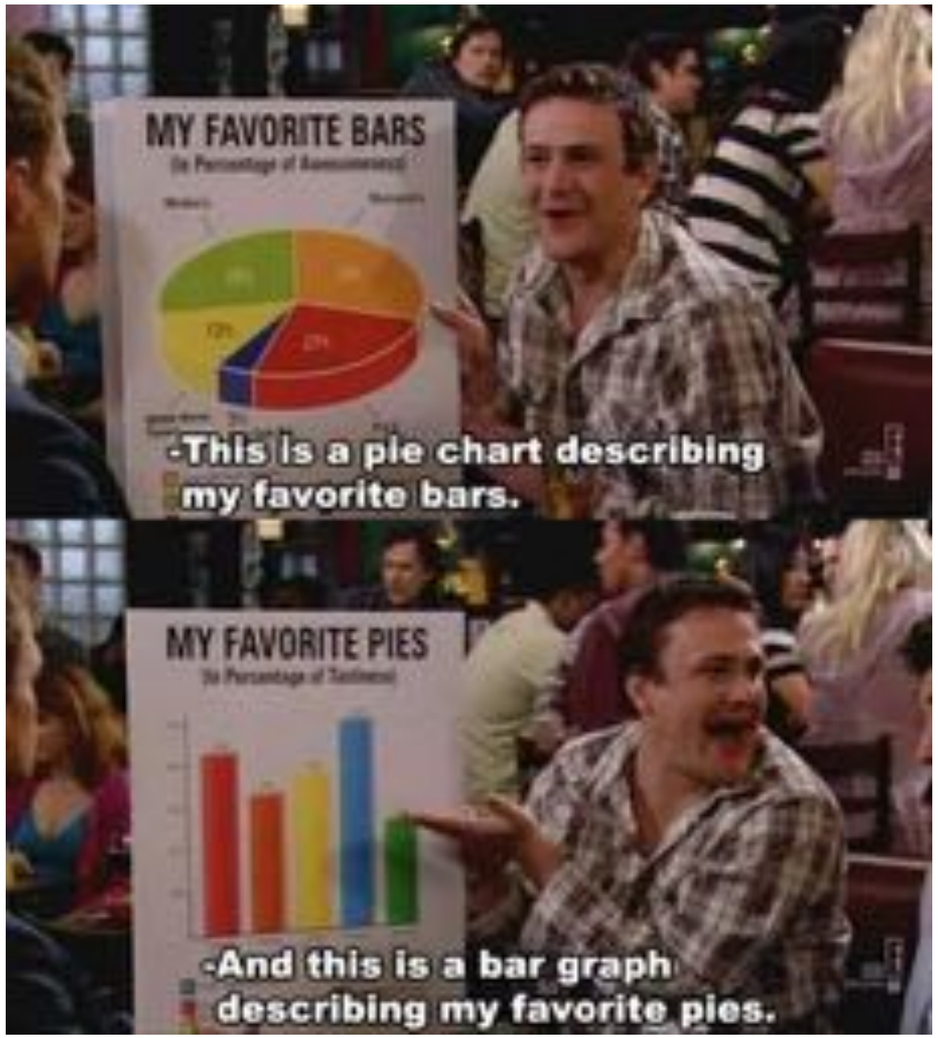
We work across **multiple verticals** on projects that range from mature, large scale implementations to proofs of concept and can provide skills in **development, architecture, delivery, training and support**.

Agenda

- Tooling
- Data Science Steps
- Demo

Tooling

Analytics!

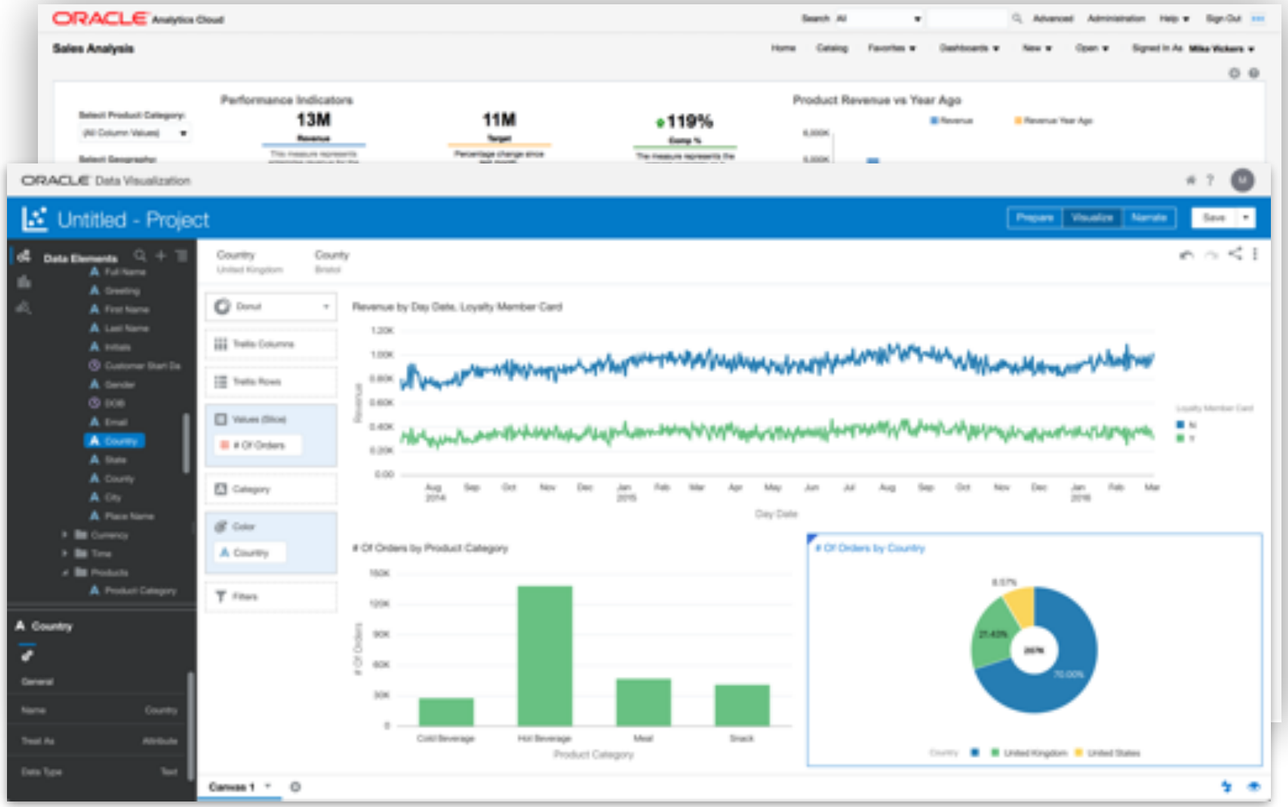

MY FAVORITE BARS
-This is a pie chart describing my favorite bars.
MY FAVORITE PIES
-And this is a bar graph describing my favorite pies.

Oracle Analytics Cloud

- Oracle's complete suite of Platform Services (PaaS) for unified analytics in the cloud
- Delivered entirely in the cloud:
  - No infrastructure footprint
  - Flexibility to scale up or down based on your immediate needs
  - Simplified, metered licensing
- Several options to suit your needs:
  - Oracle or customer/partner managed services
  - Functionality bundled into 3 editions

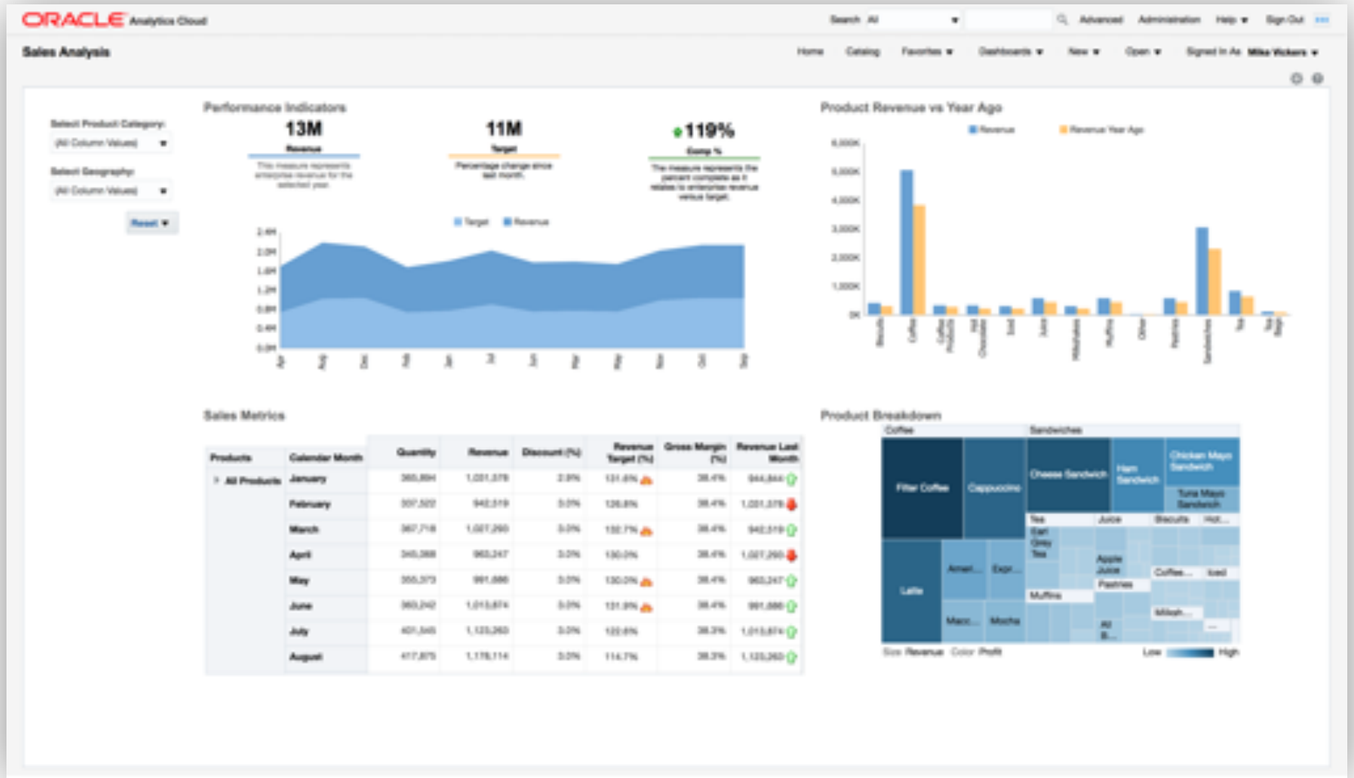
Functions

- OAC supports **every** type of analytics workload across your organisation
- *Classic* enterprise BI:
  - Analysis & dashboarding
  - Published reporting
  - Enterprise Performance Management
- *Modern* departmental/personal discovery:
  - Extended data mashup & modelling
  - Data preparation, exploration & visualisation
  - Data science & machine learning



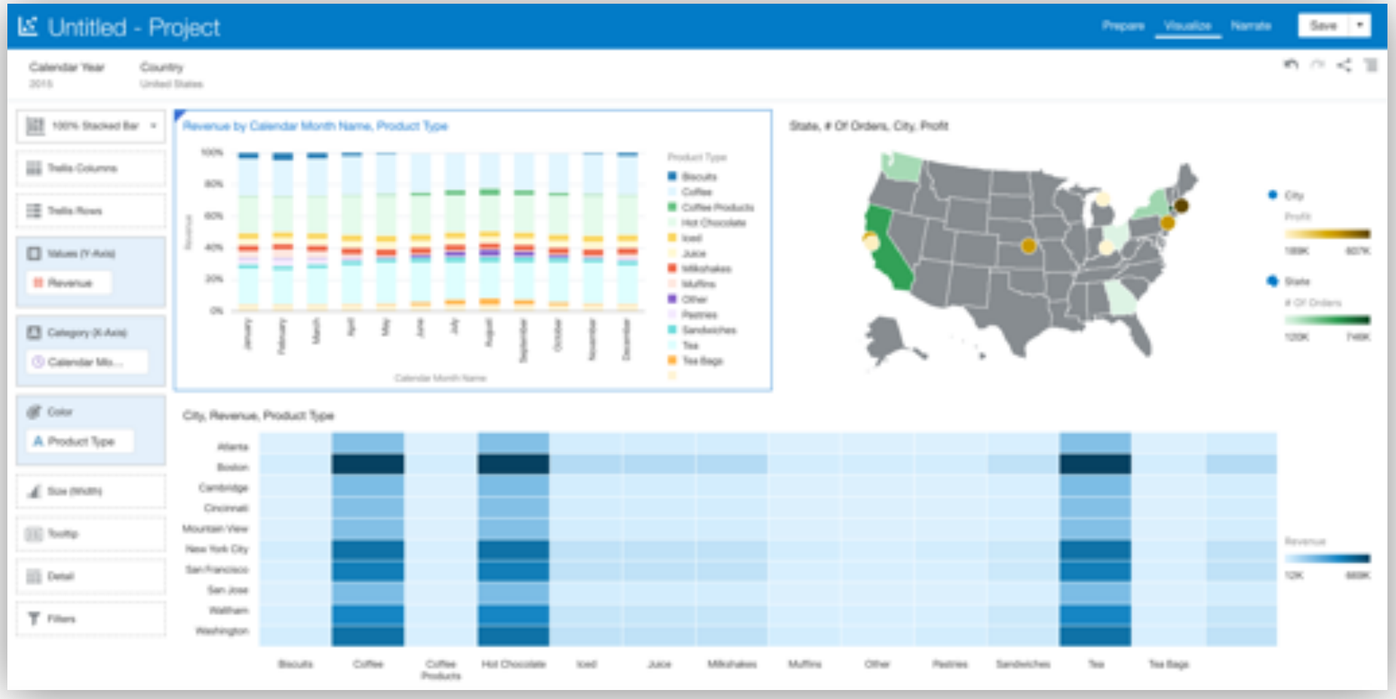
Classic Enterprise BI

- Similar User Experience to OBIEE 12c
  - Centrally maintained & governed
  - Semantic model remains key
- Interactive Dashboards
  - Ideal for 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 from multiple connections
  - Apply enrichments data prior to analysis
  - Define repeatable preparation flows
- Data Visualisation
  - Create visual insights rapidly
  - Construct narated storyboards
  - Share findings
- Machine Learning
  - Build & train ML models
  - Apply model to new data sets



Three Edition Options

Standard Edition

- Data Discovery
- Data Preparation
- What-If Planning
- Big Data Storage
- Data Transformation via Apache Spark
- Data Lake Connectivity
- Enterprise Analysis & Dashboarding
- Published Reporting
- Day by Day

Two Purchasing Options

Monthly Flex

Based on **Universal Credits** model
12 month minimum tenure
Payments made in **advance**
Unused credits are **forfeited**

Suitable for:
Predictable, production workloads
Long running platforms

Pay As You Go

Based on **Universal Credits** model
No minimum tenure
Payments made in **arrears**
Based on **consumption**

Suitable for:
Rapid Prototyping
Testing & Sampling
Elastic Scalable

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

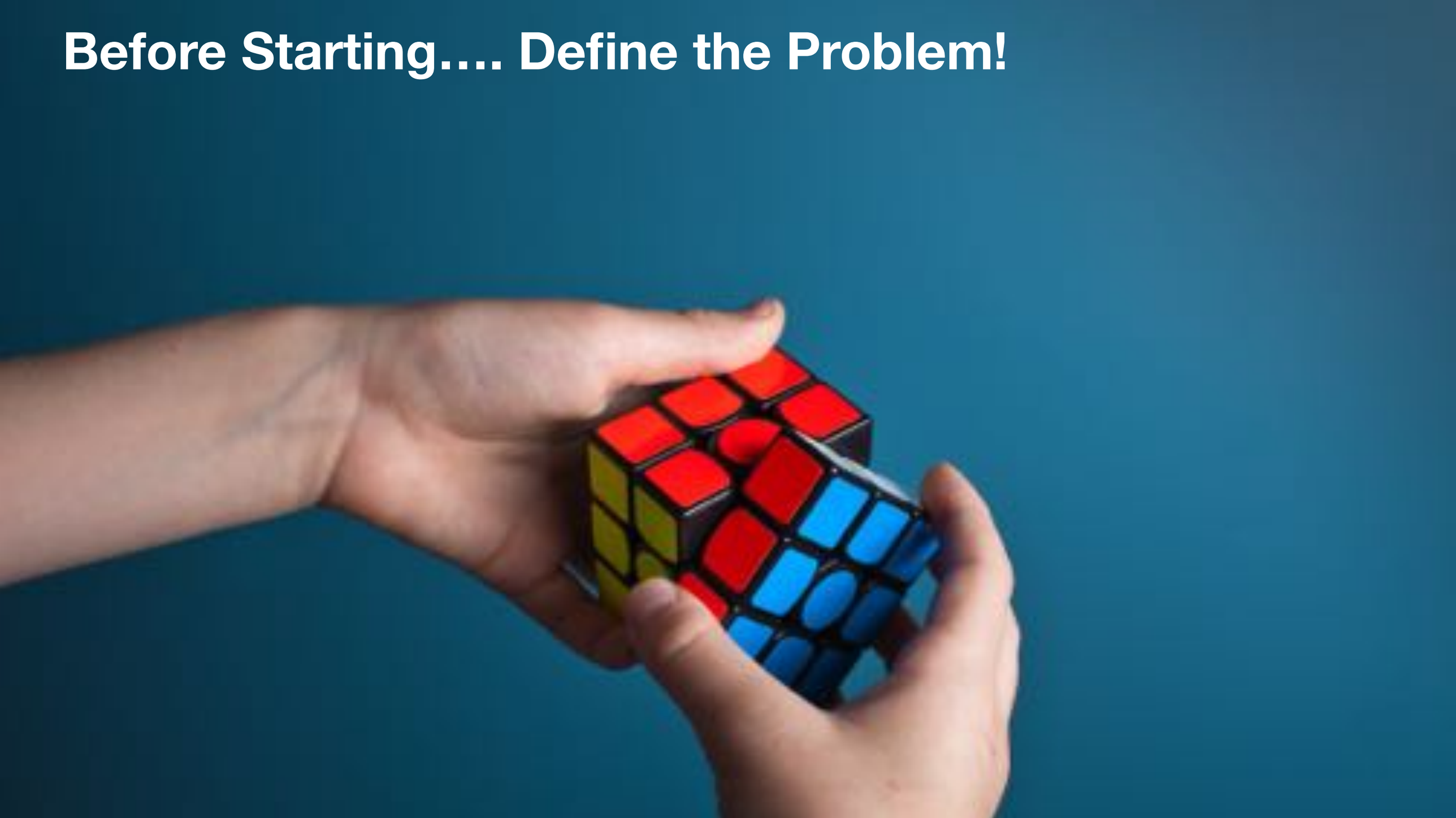
YES/NO

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 >= 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

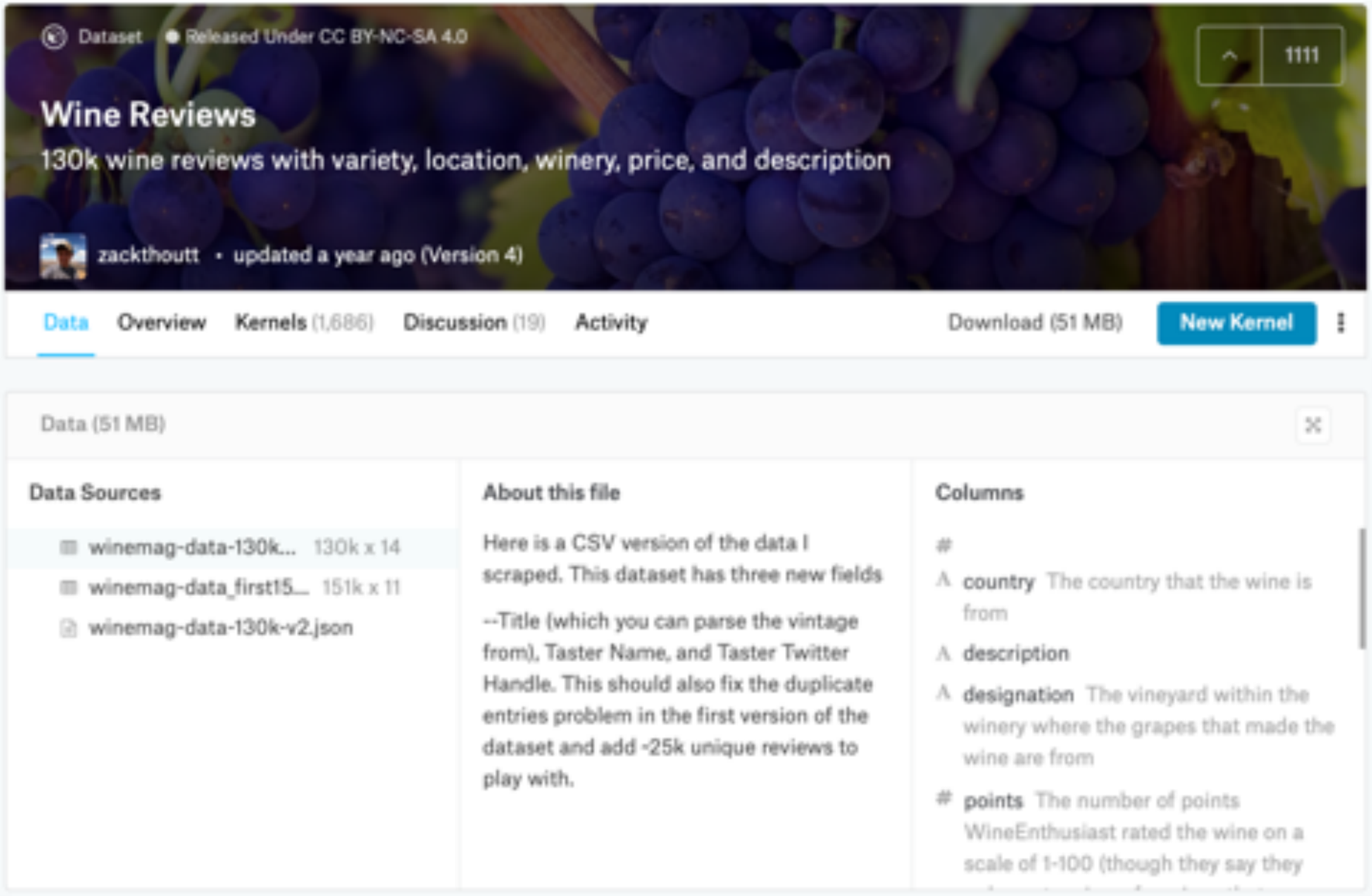
Accuracy

Accuracy

		Predicted Value	
		Good	Bad
Real Value	Good		
	Bad		

Accuracy =  / ( + )

Dataset

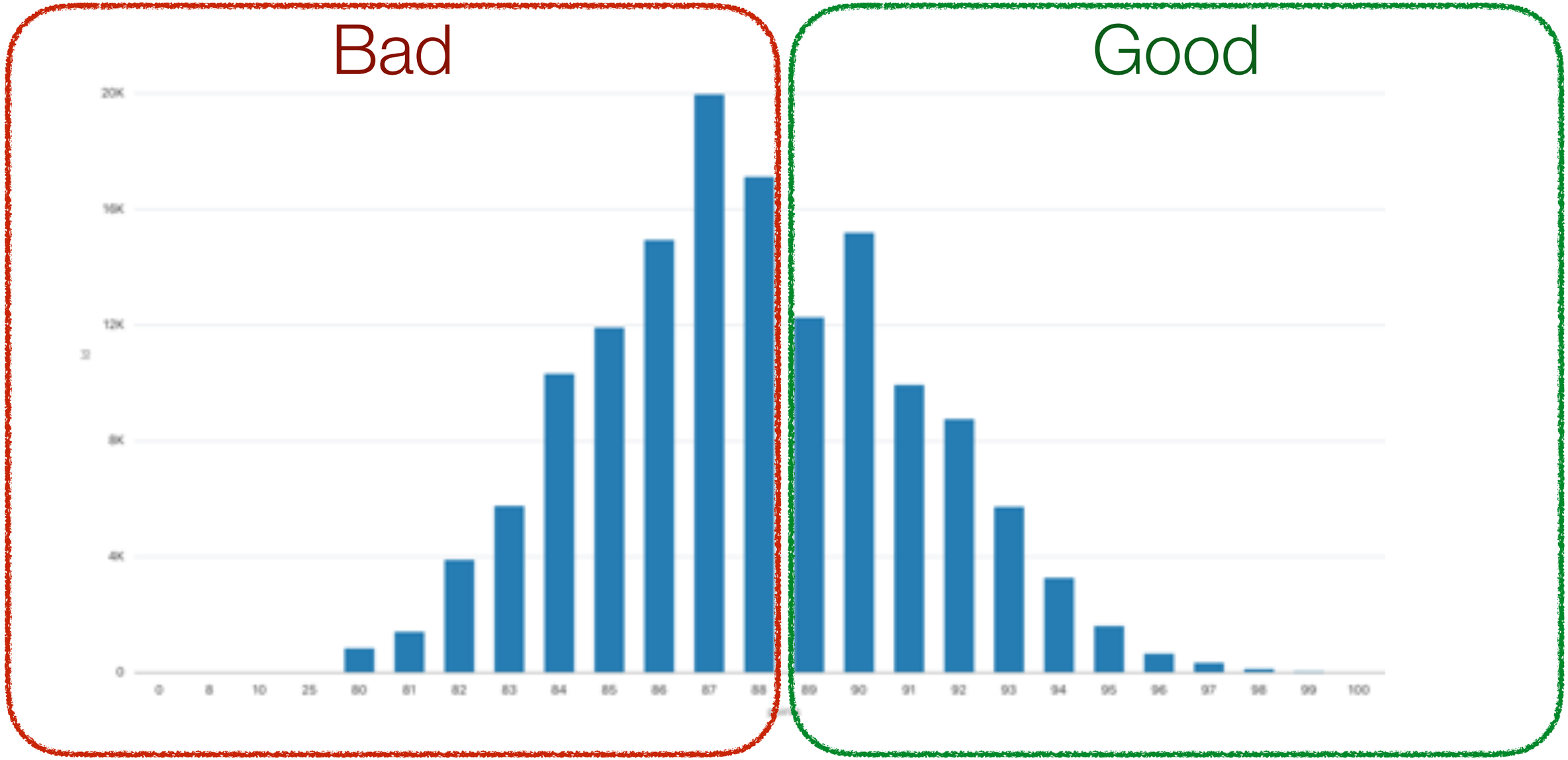


https://www.kaggle.com/zynicide/wine-reviews

The Data

A	B	C	D	E	F	G	H	I	J	K
Id	country	description	designation	points	rice	province	region_1	region_2	variety	winery
0	US	This tremendous 100%	Martha's Vineyard	96	235	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz
1	Spain	Ripe aromas of fig, blac	Carodorum Selecci	96	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez
2	US	Mac Watson honors th	Special Selected Late	96	90	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley
3	US	This spent 20 months	Reserve	96	65	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi
4	France	This is the top wine fr	La Brï¿½lade	95	66	Provence	Bandol		Provence red blend	Domaine de la Bï¿½gude
5	Spain	Deep, dense and pure	Numanthia	95	73	Northern Spain	Toro		Tinta de Toro	Numanthia
6	Spain	Slightly gritty black-fr	San Roman	95	65	Northern Spain	Toro		Tinta de Toro	Maurodos
7	Spain	Lush cedary black-frui	Carodorum Vinico G	95	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez
8	US	This re-named vineyar	Silice	95	65	Oregon	Chehalem Mountains	Willamette Valley	Pinot Noir	Bergström
9	US	The producer sources	Gap's Crown Vineyard	95	60	California	Sonoma Coast	Sonoma	Pinot Noir	Blue Farm
10	Italy	Elegance, complexity	Ronco della Chiesa	95	80	Northeastern Italy	Collio		Friulano	Borgo del Tiglio
11	US	From 18-year-old vine	Estate Vineyard Wad	95	48	Oregon	Ribbon Ridge	Willamette Valley	Pinot Noir	Patricia Green Cellars
12	US	A standout even in thi	Weber Vineyard	95	48	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Patricia Green Cellars
13	France	This wine is in peak co	Chï¿½teau Montus Pr	95	90	Southwest France	Madiran		Tannat	Vignobles Brumont
14	US	With its sophisticated	Grace Vineyard	95	185	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Domaine Serene
15	US	First made in 2006, th	Sigrid	95	90	Oregon	Willamette Valley	Willamette Valley	Chardonnay	Bergström
16	US	This blockbuster, pow	Rainin Vineyard	95	325	California	Diamond Mountain D	Napa	Cabernet Sauvignon	Hall
17	Spain	Nicely oaked blackber	6 Avitos Reserva Pre	95	80	Northern Spain	Ribera del Duero		Tempranillo	Valduero
18	France	Coming from a seven-	Le Pigeonnier	95	290	Southwest France	Cahors		Malbec	Chï¿½teau Lagrï¿½ze
19	US	This fresh and lively m	Gap's Crown Vineyard	95	75	California	Sonoma Coast	Sonoma	Pinot Noir	Gary Farrell

Good/Bad


Bad
Good
20K
16K
12K
8K
4K
0
0
8
10
25
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

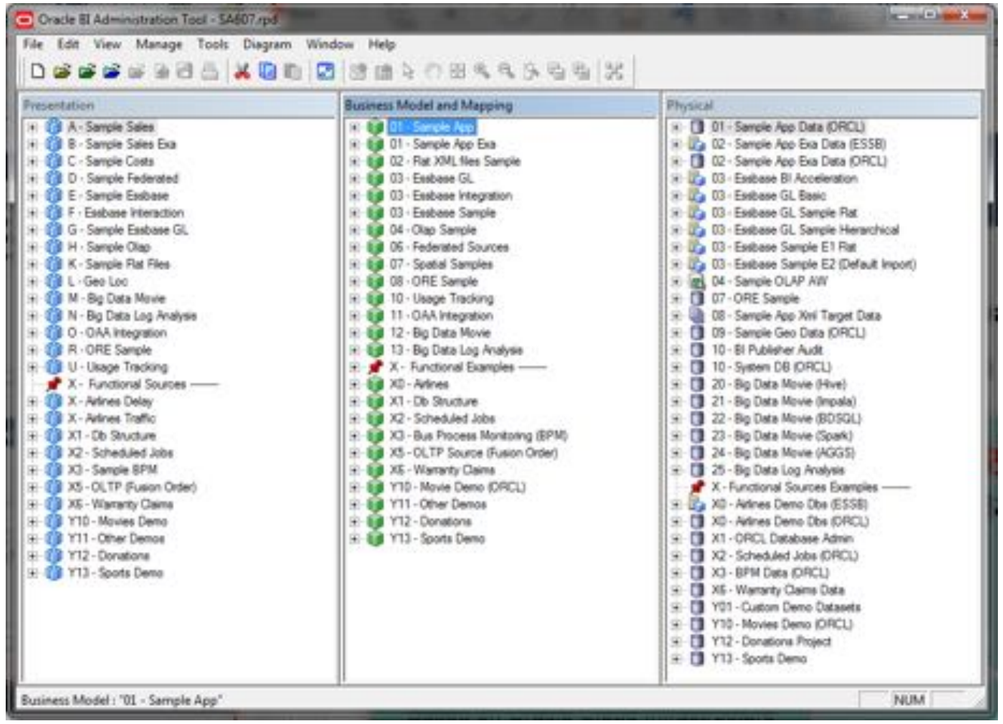
Become a Data Scientist with OAC

Connect	Clean	Transform & Enrich
Analyse	Train & Evaluate	Predict

Connect

Data Sources

Pre-Defined
Data Models



Clean

N/A

Missing Values

Mark <> MArk

Wrong Values

City
“Rome”

Irrelevant Observations

Col1 -> Name

Labelling Columns

Role: CIO
Salary:500 K$

Handling Outliers

0-200k
0-1

Feature Scaling

Of Clicks

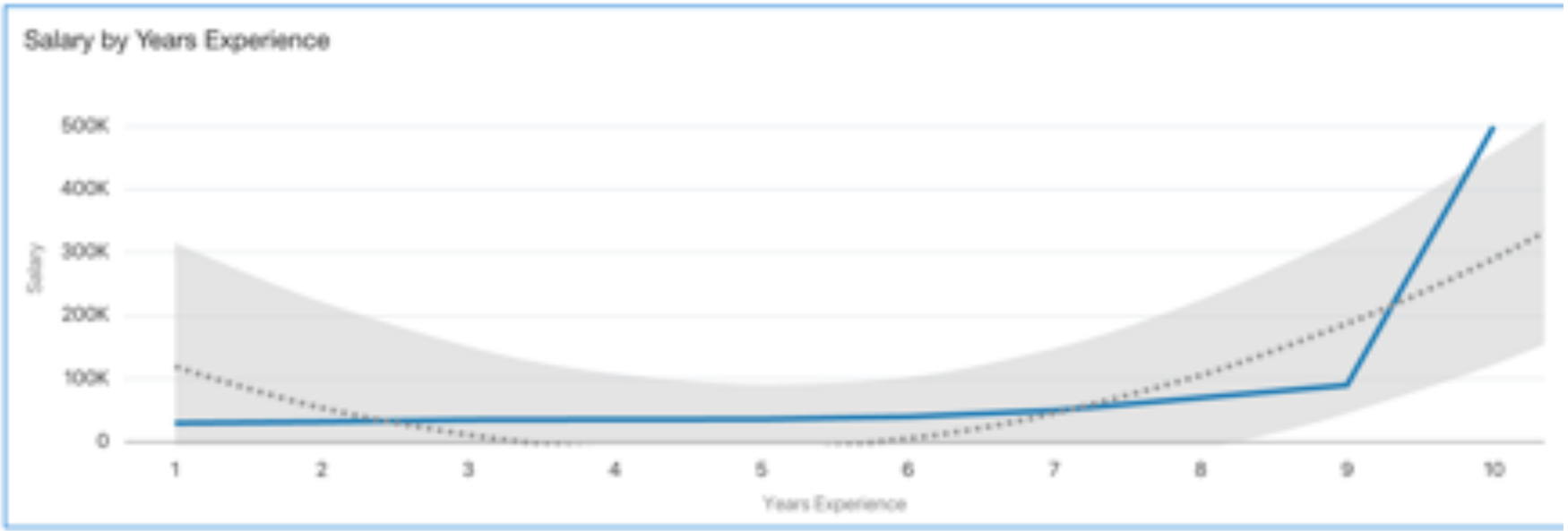
Aggregation

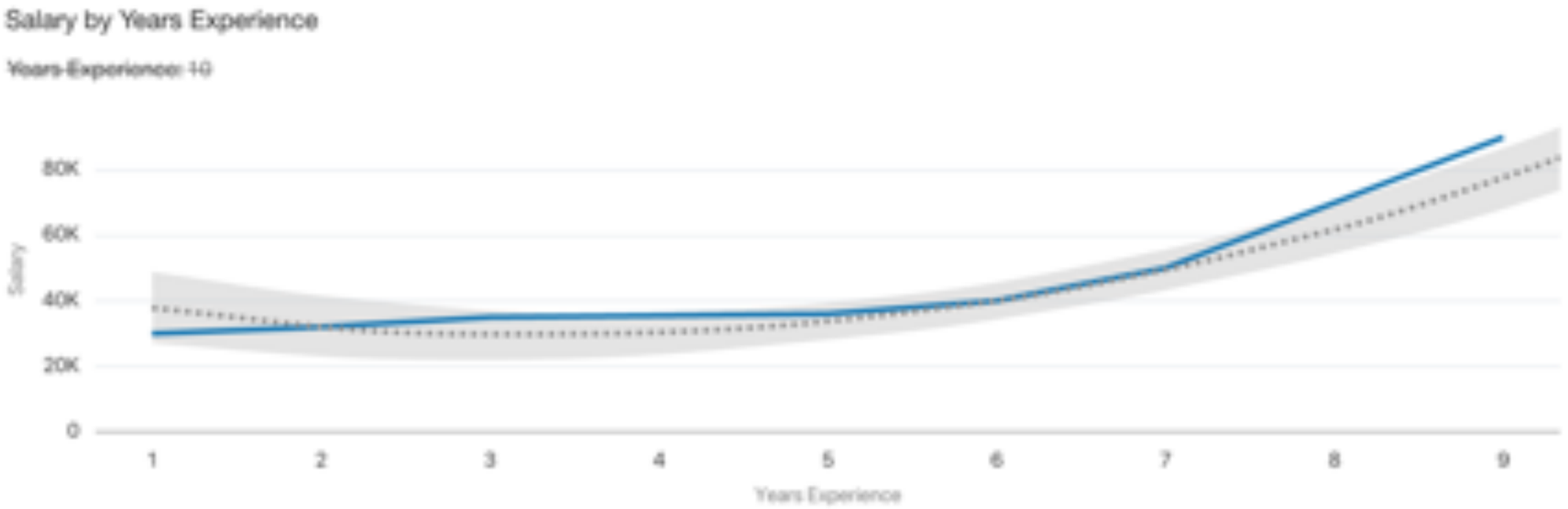
Train: 80%
Test: 20%

Train/Test Set Split

Why Removing an Outlier?

Years Experience	Salary
1	30.000
2	32.000
3	35.000
4	35.500
5	36.000
6	40.000
7	50.000
8	70.000
9	90.000
10	500.000





Enrich - Feature Engineering

Location -> ZIP Code

Additional Data Sources?

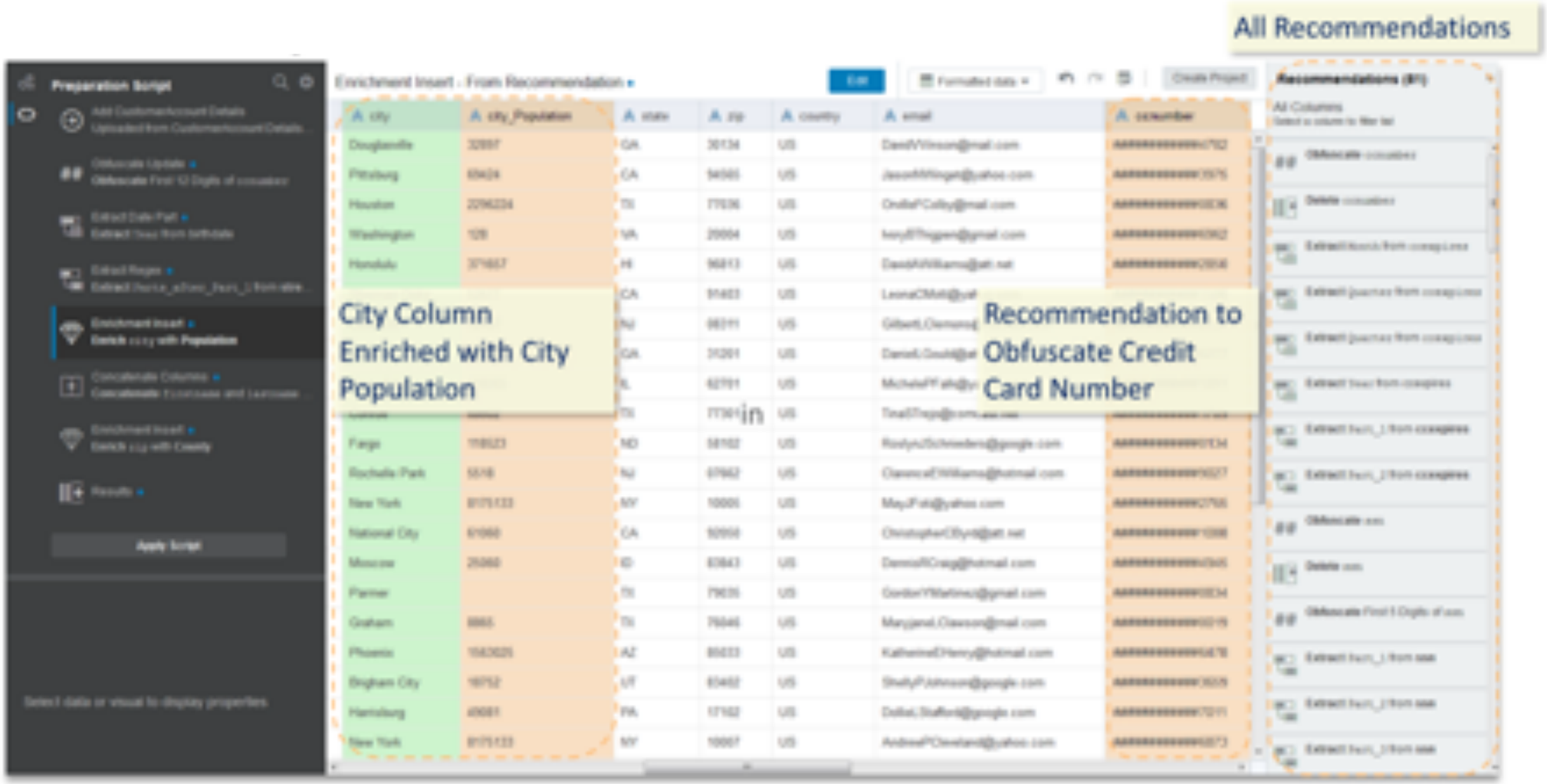
Name -> Sex

2 Locations -> Distance

Data Flow

Day/Month/Year -> Date

Data Preparation Recommendations



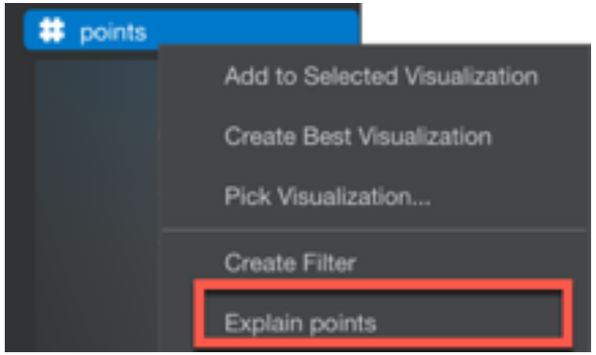
Analyse - Data Overview

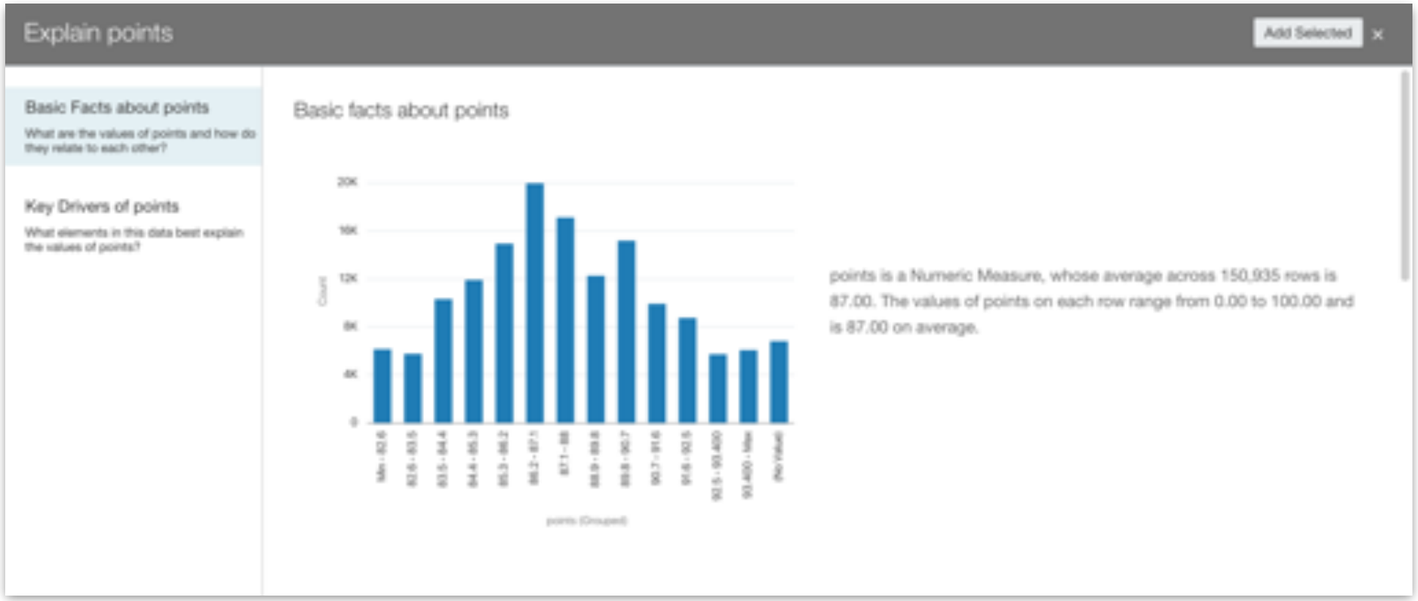
Results

Metadata

Data Element	Data Type	Treat As	Aggregation	Sample Values
Id	varchar(80)	Attribute	none	1470; 817; 1028; 632; 3689; 4148; 2576; 963; 4979; 281
country	varchar(137)	Attribute	none	US; France; Italy; Spain; Portugal; Germany; Argentina; Chile; Austria; Greece
country_continent	varchar(4000)	Attribute	none	NA; EU
country_fips	varchar(4000)	Attribute	none	US; IT; FR
country_iso3	varchar(4000)	Attribute	none	USA; FRA; ITA
country_iso_numeric	number	Measure	sum	840; 380; 250
country_iso2	varchar(4000)	Attribute	none	US; IT; FR
description	varchar(1247)	Attribute	none	This elegant wine combines subtle nutmeg and cardamom aromas with crisp app....
designation	varchar(122)	Attribute	none	Reserve; Estate; Reserva; Riserva; Estate Bottled; Vieilles Vignes; Crianza; Classic...
points	number	Measure	sum	90; 89; 88; 87; 91; 86; 92; 93; 85; 94
price	varchar(15)	Attribute	none	25; 20; 40; 18; 60; 30; 28; 35; 50; 15
province	varchar(53)	Attribute	none	California; Oregon; Bordeaux; Tuscany; Piedmont; Washington; Northern Spain; M...
region_1	varchar(75)	Attribute	none	Willamette Valley; Napa Valley; Barolo; Brunello di Montalcino; Russian River Valle...
region_2	varchar(35)	Attribute	none	Central Coast; Sonoma; Willamette Valley; Napa; Columbia Valley; Mendocino/La...
variety	varchar(53)	Attribute	none	Pinot Noir; Chardonnay; Bordeaux-style Red Blend; Cabernet Sauvignon; Red Ble...
winery	varchar(84)	Attribute	none	Tarara; Heron Hill; Byron; Bergstr√∂m; Herdade do Rocim; Rusack; Sarah's Viney...

Analyse - Explain



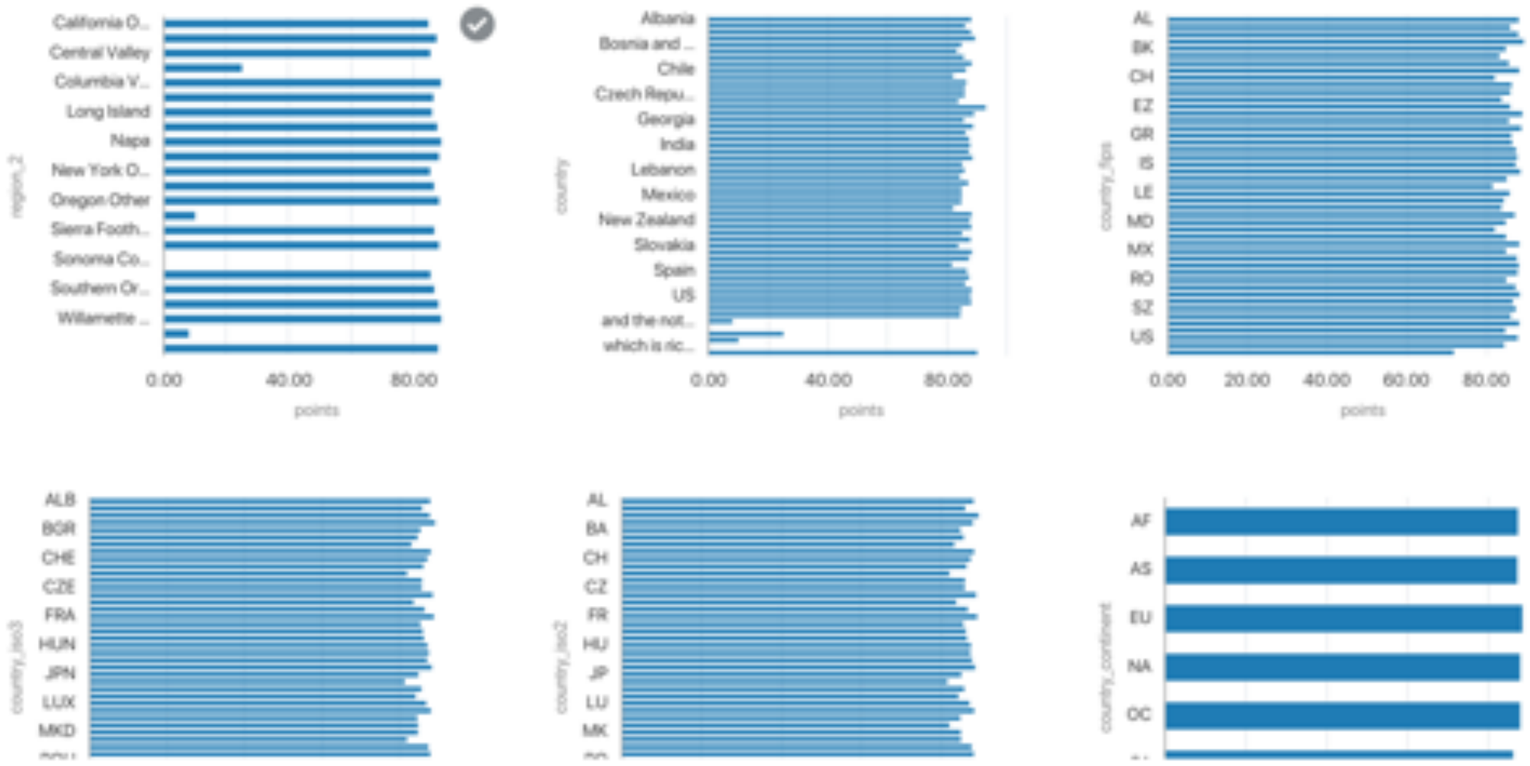


Explain - Key Drivers

Key Drivers of points

The 6 attributes most strongly correlated to outcomes for points are: **region_2, country, country_fips, country_iso3, country_iso2, country_continent**

The charts below show the distribution of points values across each of the key drivers. Click the checkmarks above any of the visuals to add them to your project when done.

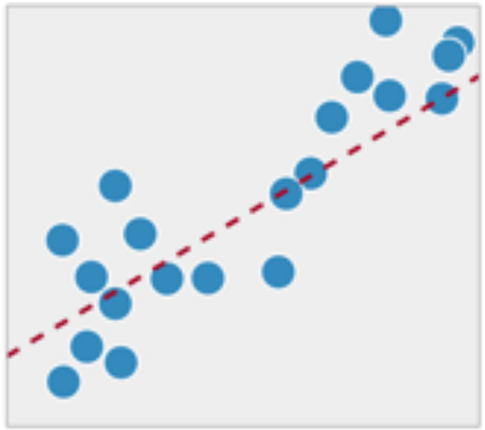


Train - What Problem are we Trying to Solve?

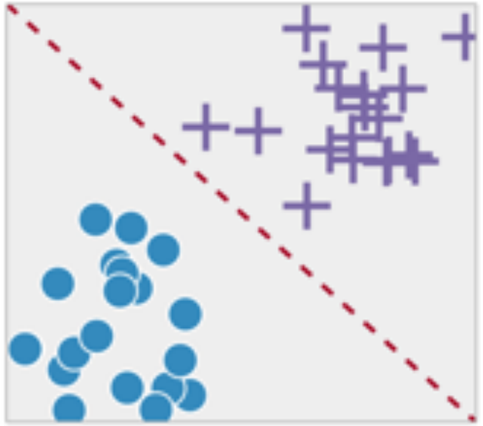
Supervised

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

Regression



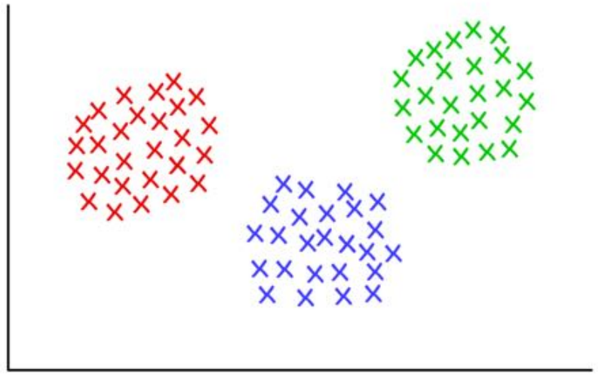
Classification



Unsupervised

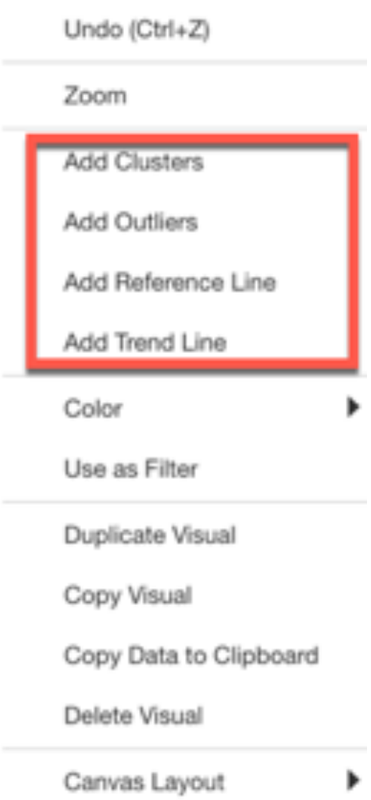
"Here is a dataset, make sense out of it!"

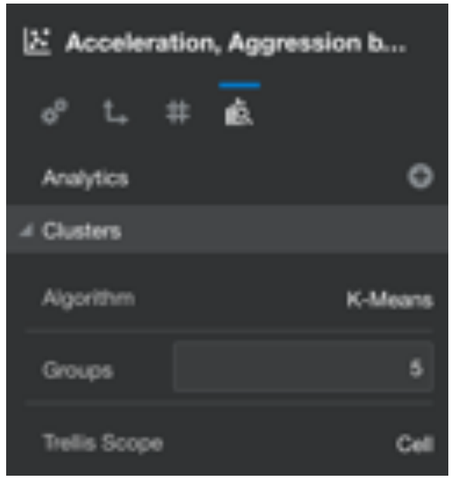
Clustering

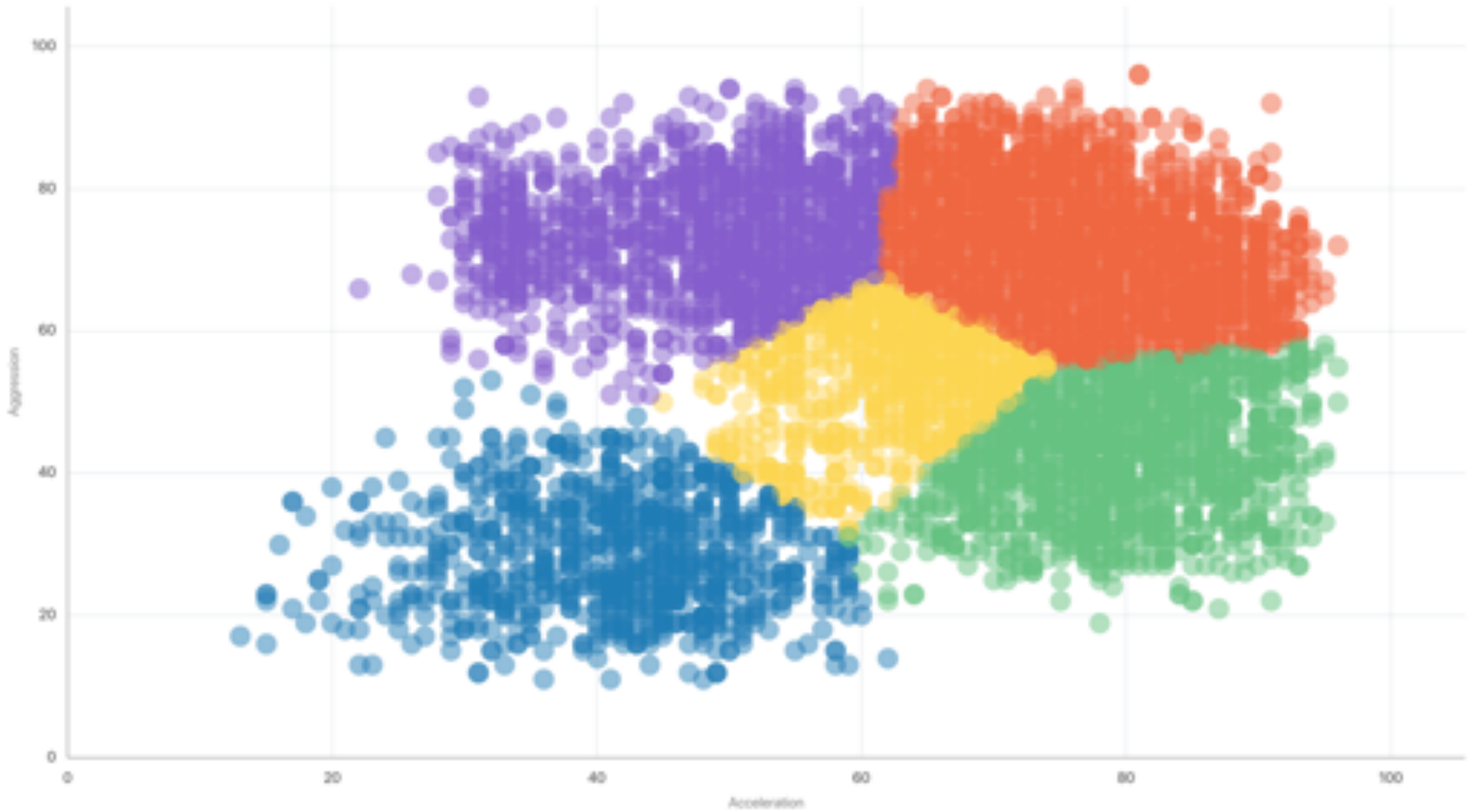


https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

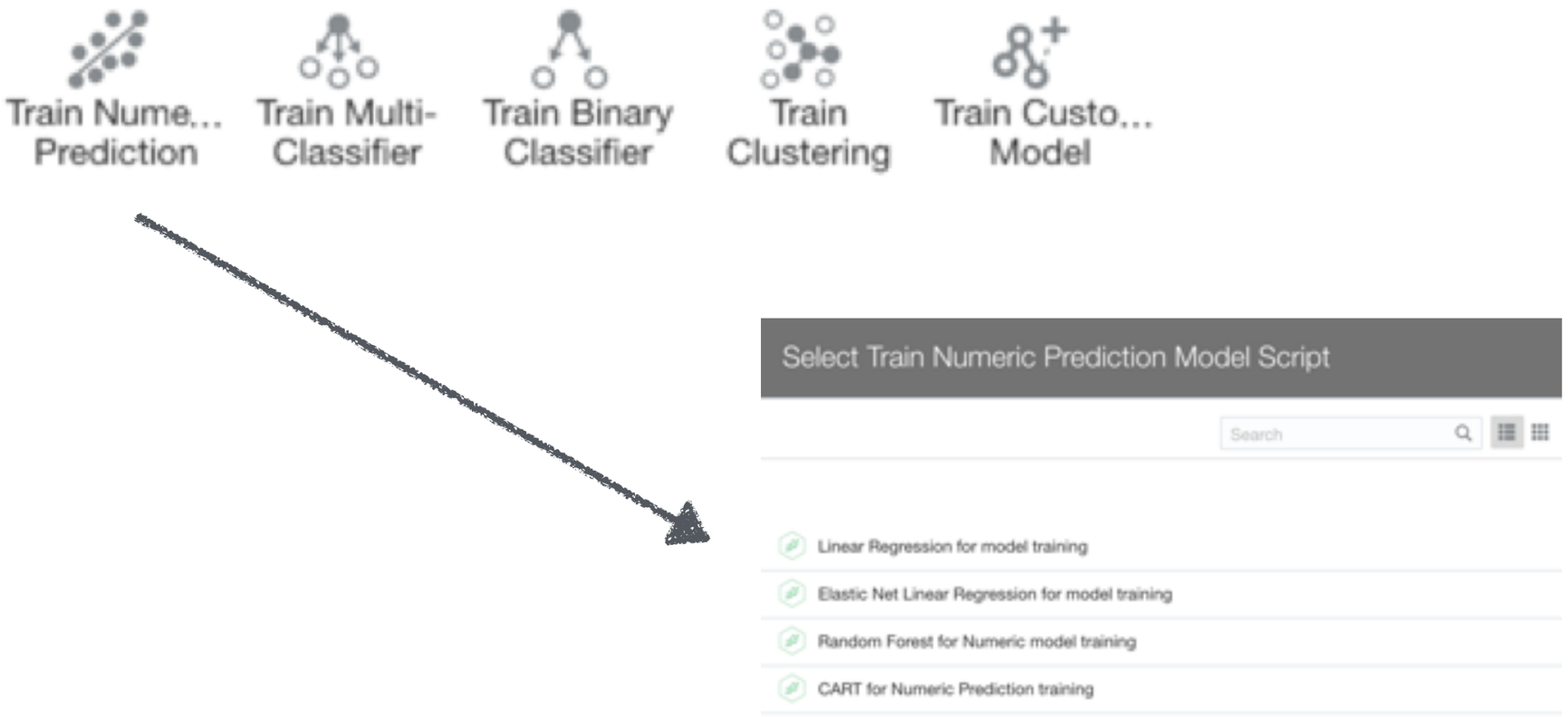
Easy Models


Undo (Ctrl+Z)
Zoom
Add Clusters
Add Outliers
Add Reference Line
Add Trend Line
Color
Use as Filter
Duplicate Visual
Copy Visual
Copy Data to Clipboard
Delete Visual
Canvas Layout

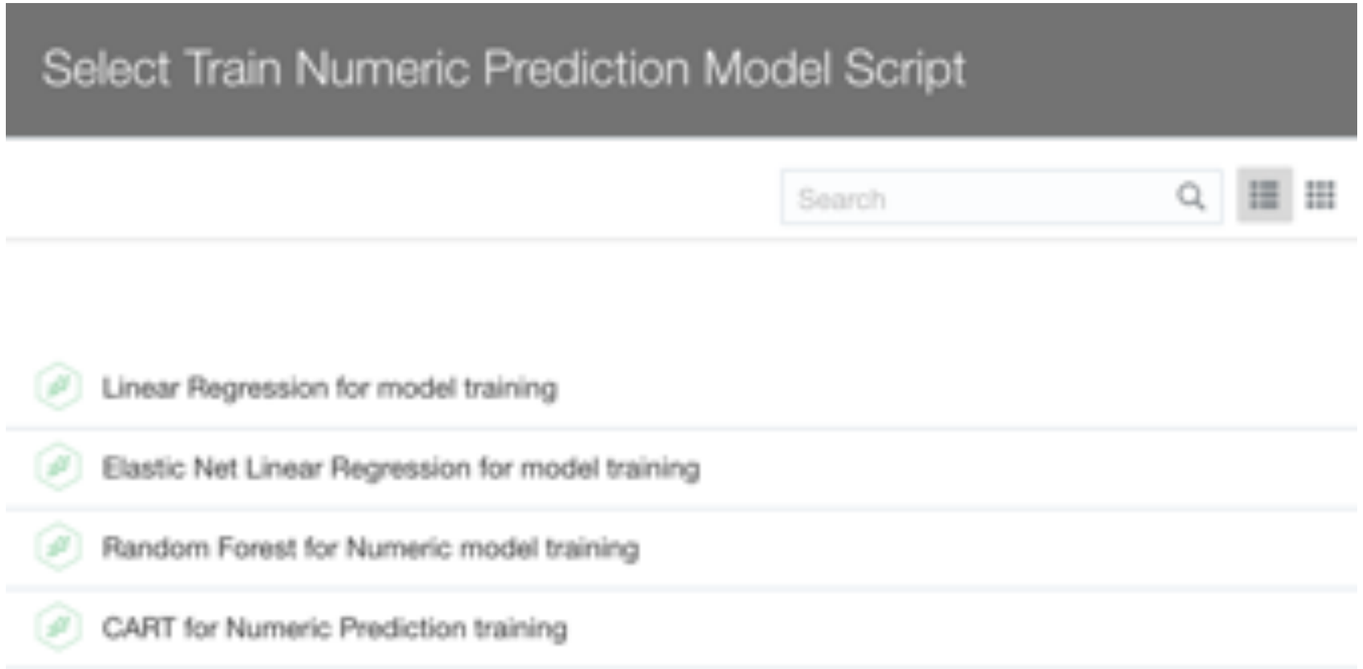

Acceleration, Aggression b...
Analytics
Clusters
Algorithm
K-Means
Groups
5
Trellis Scope
Cell

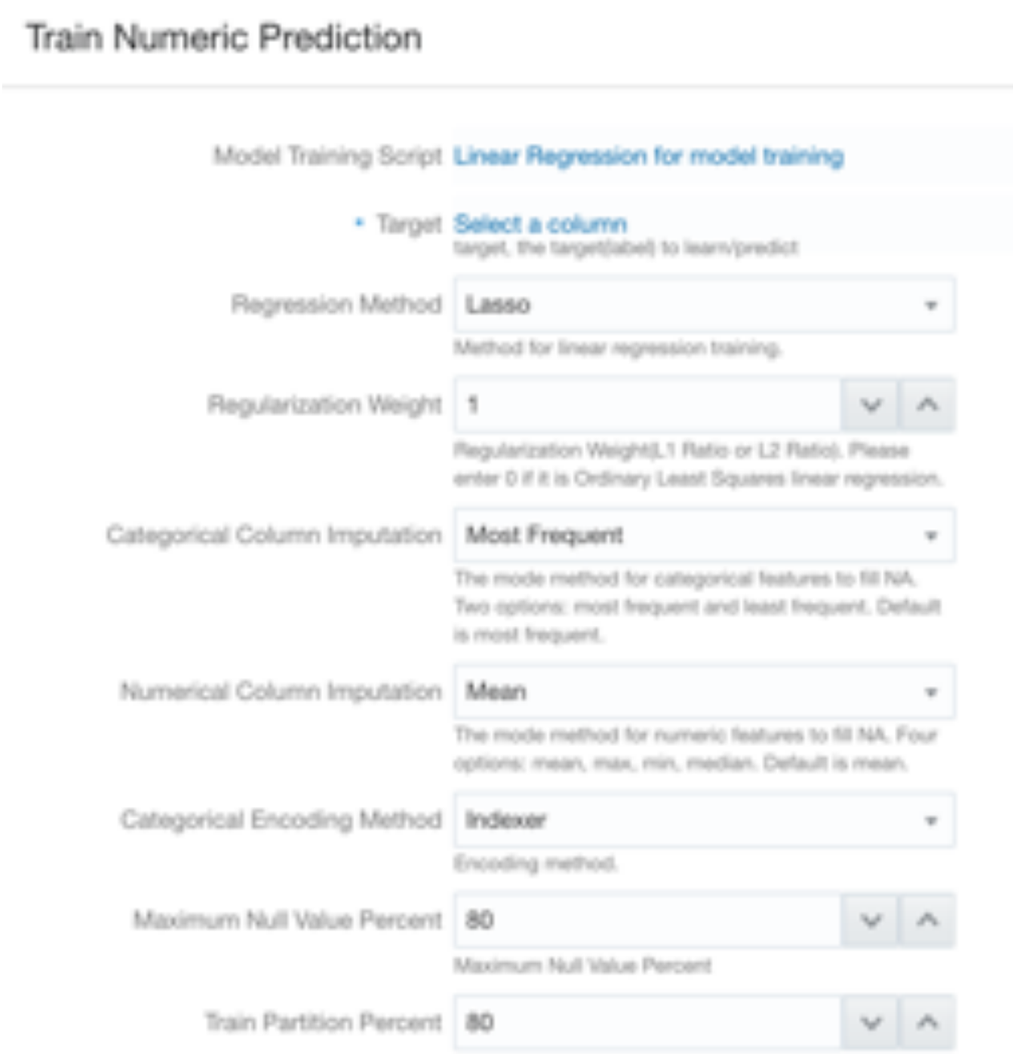

100
80
60
40
20
0
Aggression
0
20
40
60
80
100
Acceleration

DataFlow Train Model



Which Model - Parameters To Pick?





Select, Try, Save, Change, Try, Save …..

Train Numeric Prediction

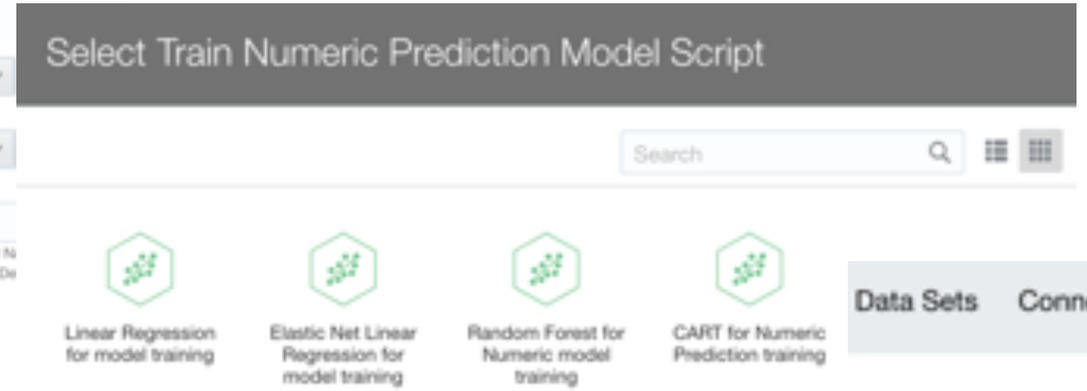
Model Training Script Elastic Net Linear Regression for model training

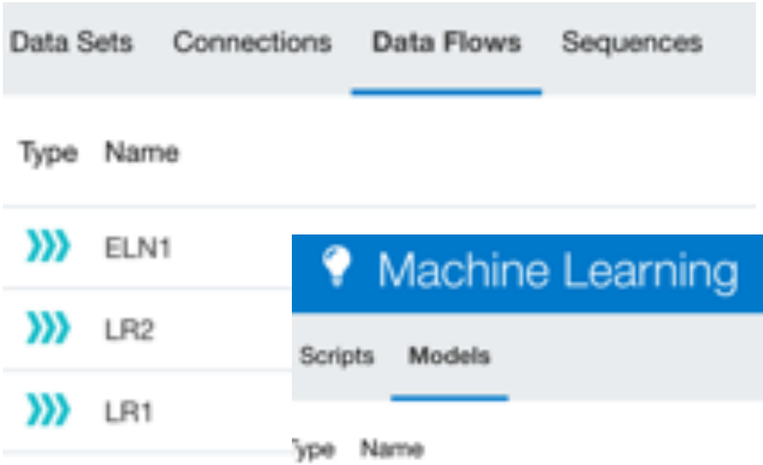
* Target points
target, the target(label) to learn/predict

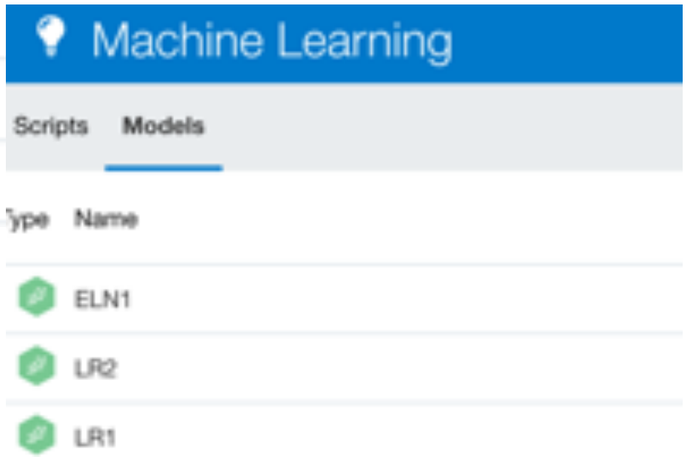
L1 Ratio 0.5
L1 Ratio

L2 Ratio 0.5
L2 Ratio

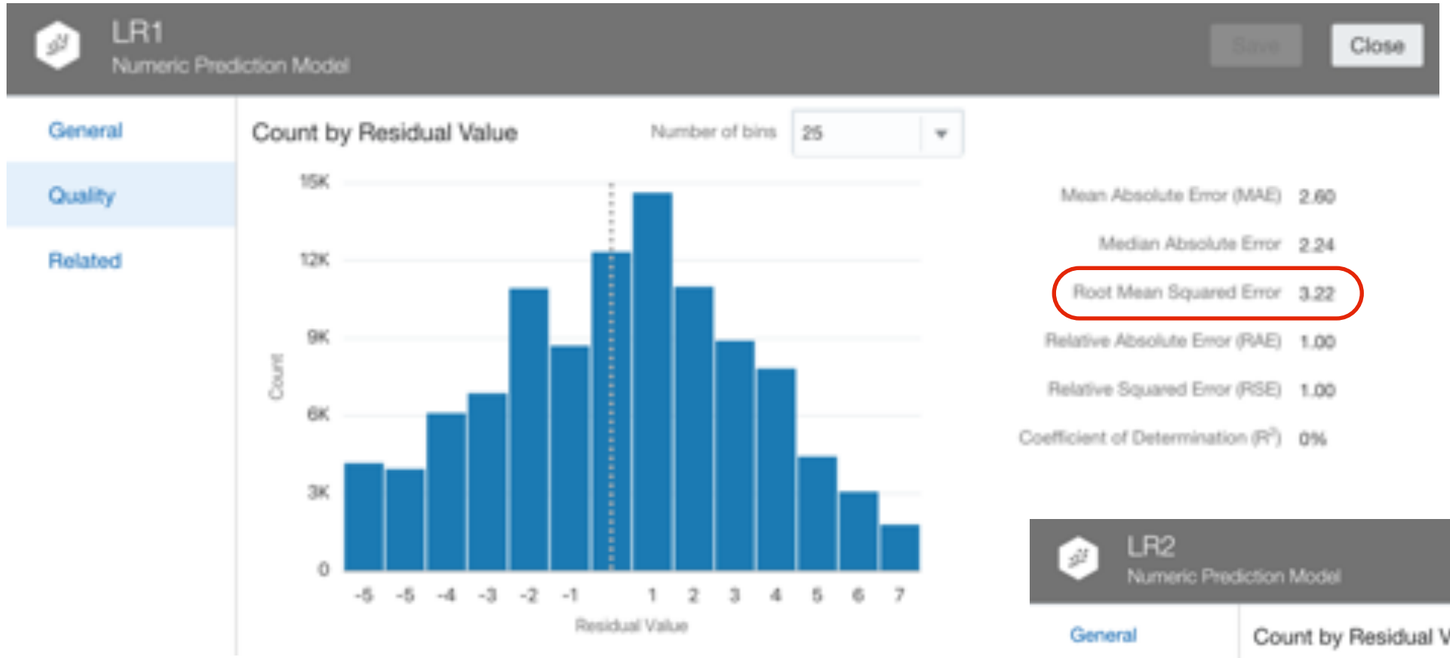
Categorical Column Imputation Most Frequent
The mode method for categorical features to fill N
Two options: most frequent and least frequent. De
is most frequent.

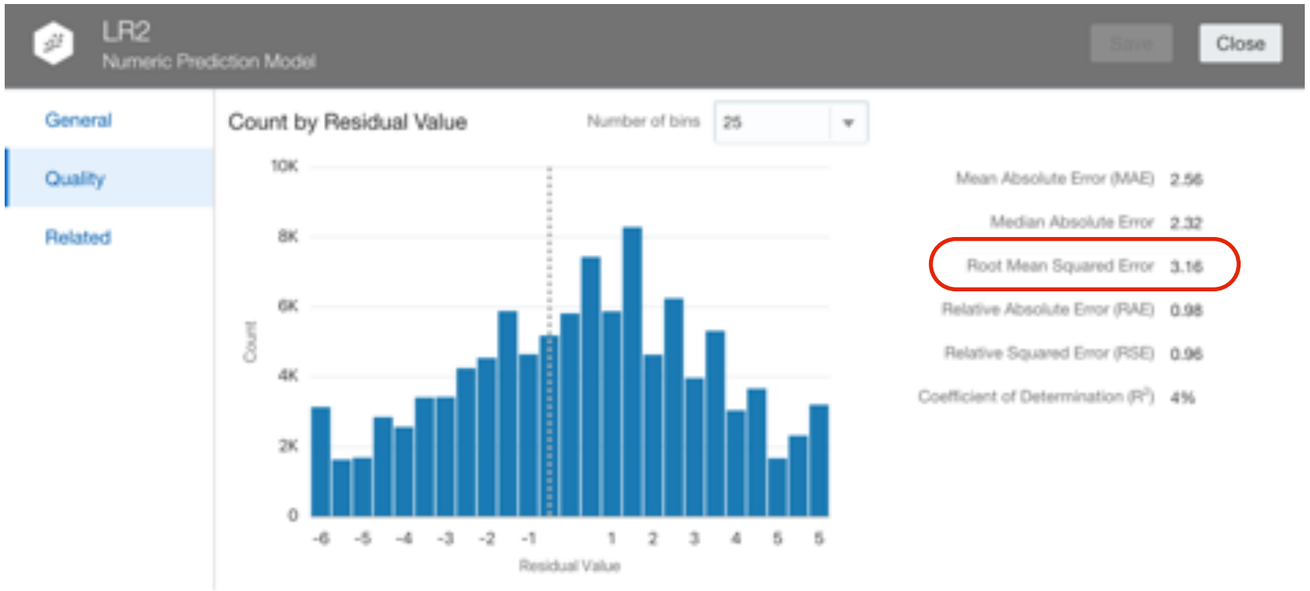






Compare


LR1
Numeric Prediction Model
Save
Close
General
Quality
Related
Count by Residual Value
Number of bins
25
15K
12K
9K
6K
3K
0
Count
-5 -5 -4 -3 -2 -1 1 2 3 4 5 6 7
Residual Value
Mean Absolute Error (MAE) 2.60
Median Absolute Error 2.24
Root Mean Squared Error 3.22
Relative Absolute Error (RAE) 1.00
Relative Squared Error (RSE) 1.00
Coefficient of Determination (R²) 0%


LR2
Numeric Prediction Model
Save
Close
General
Quality
Related
Count by Residual Value
Number of bins
25
10K
8K
6K
4K
2K
0
Count
-6 -5 -4 -3 -2 -1 1 2 3 4 5 5
Residual Value
Mean Absolute Error (MAE) 2.56
Median Absolute Error 2.32
Root Mean Squared Error 3.16
Relative Absolute Error (RAE) 0.98
Relative Squared Error (RSE) 0.96
Coefficient of Determination (R²) 4%

Compare

Predicted Values

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

There is No Single Truth…

Predicted Values

Actual Values	0.0	1.0	Total
0.0	40408	502	40910 (90%)
1.0	3731	896	4627 (10%)
Total	44139 (97%)	1398 (3%)	45537 (100%)

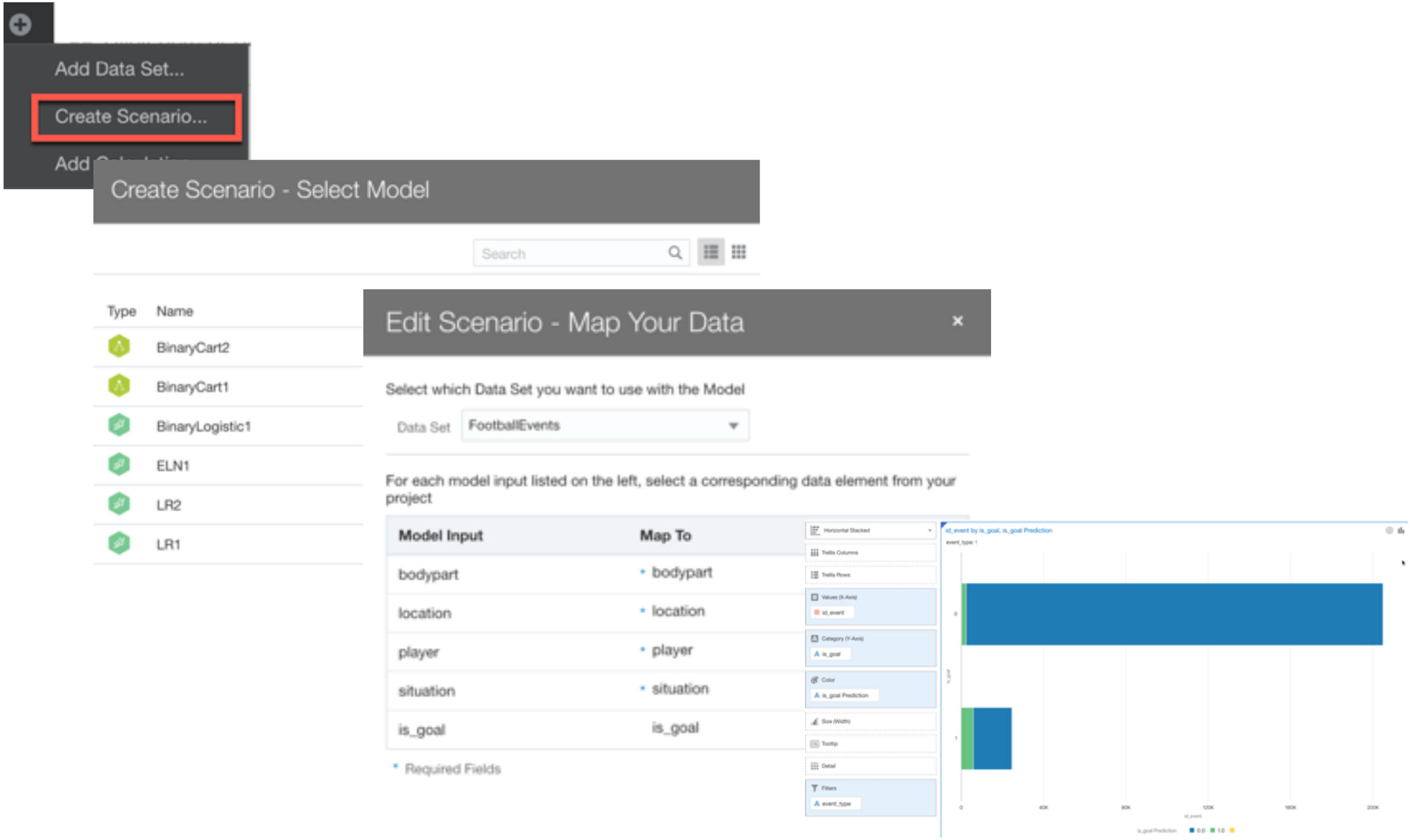
502/(502+896) = **64.09%**

471/(471+866)=**64.77%**

Predicted Values

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

Predict - Use On the Fly



Predict - Step of a Data Flow

Apply Model

Select Model

Search

Type	Name	Outputs	Modified
	BinaryCart2	is_goal	54 minutes ago
	BinaryCart1	is_goal	51 minutes ago
	BinaryLogistic1	is_goal	12:25 PM
	ELN1	points	11:56 AM
	LR2	points	11:36 AM
	LR1	points	10:31 AM

Model BinaryCart2

Outputs

Create	Output	Column Name
✓	PredictedValue	PredictedValue
✓	PredictionConfidencePercentage	PredictionConfidencePercentage
✓	PredictionGroup	PredictionGroup

Inputs

Model	Input
situation	situation
bodypart	bodypart
player	player

Demo

Conclusions

73% > 63% > 50%

Data Cleaning & Transformation

Model Creation & Evaluation

Trial Error

Visual -> UI Driven

Existing Skillset

Example

Wine Ratings Prediction using Machine Learning

Olivier Goutay Follow

Jun 14, 2018 · 5 min read

Image from wall2born.com

https://towardsdatascience.com/wine-ratings-prediction-using-machine-learning-ce259832b321

Custom ML Model

Available only in OAC Classic


A MEAL
WITHOUT WINE
IS CALLED
BREAKFAST

Is It Corked? Wine Machine Learning Predictions with OAC

Francesco Tisiot
BI Tech Lead at Rittman Mead

ODTUG Kscope19

SEATTLE, WASHINGTON • JUNE 23-27

PLEASE FILL OUT YOUR EVALUATIONS

SEATTLE

Washington State Convention Center