

Become a Data Scientist

Francesco Tisiot
Analytics Tech Lead

rittmanmead
A DATA AND ANALYTICS COMPANY





Francesco Tisiot

Analytics Tech Lead



Verona, Italy



<http://ritt.md/ftisiot>



Over10 Years in Analytics



ft@rittmanmead.com



@FTisiot

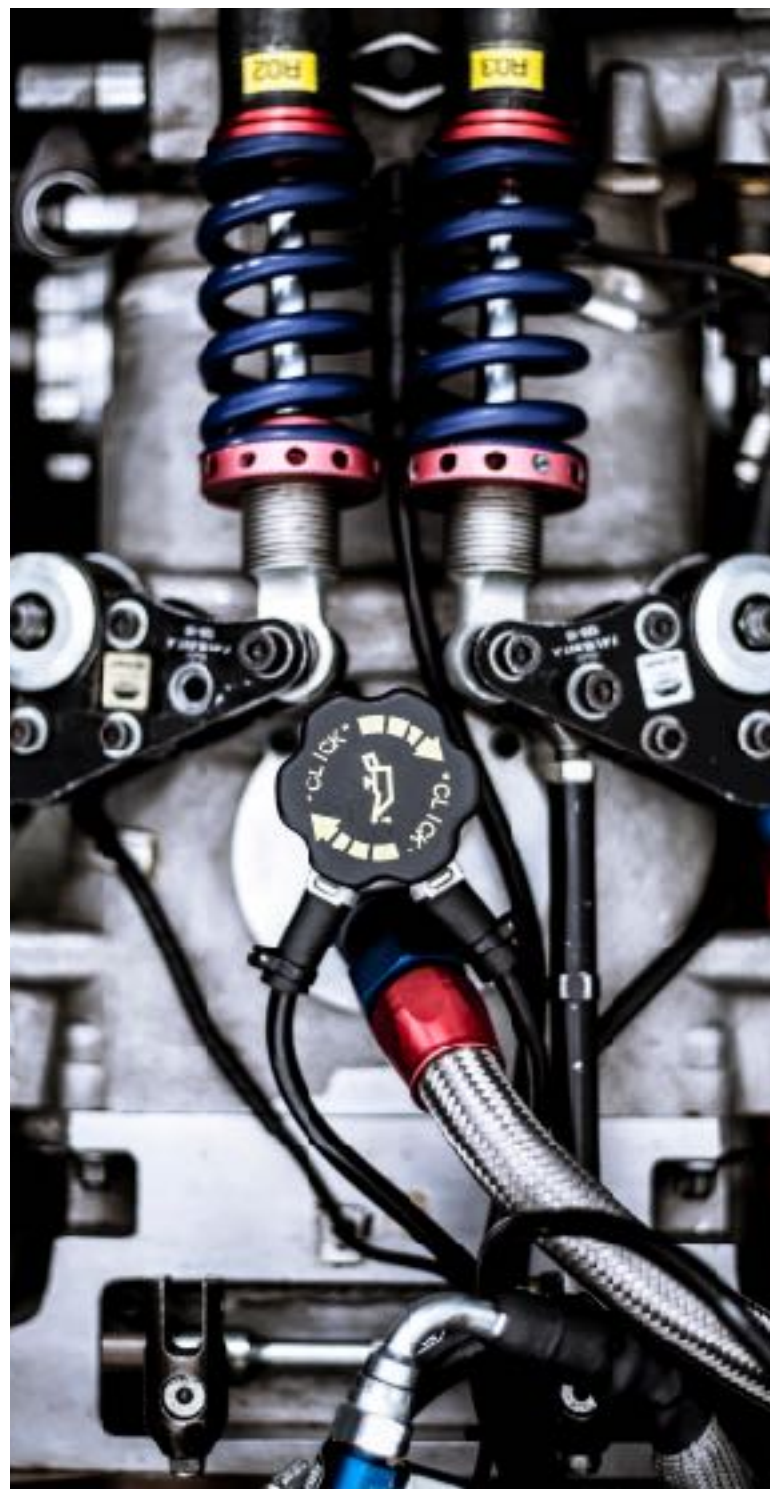


Oracle ACE Director



ITOUG Board President





Data Engineering



Analytics



Data Science

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

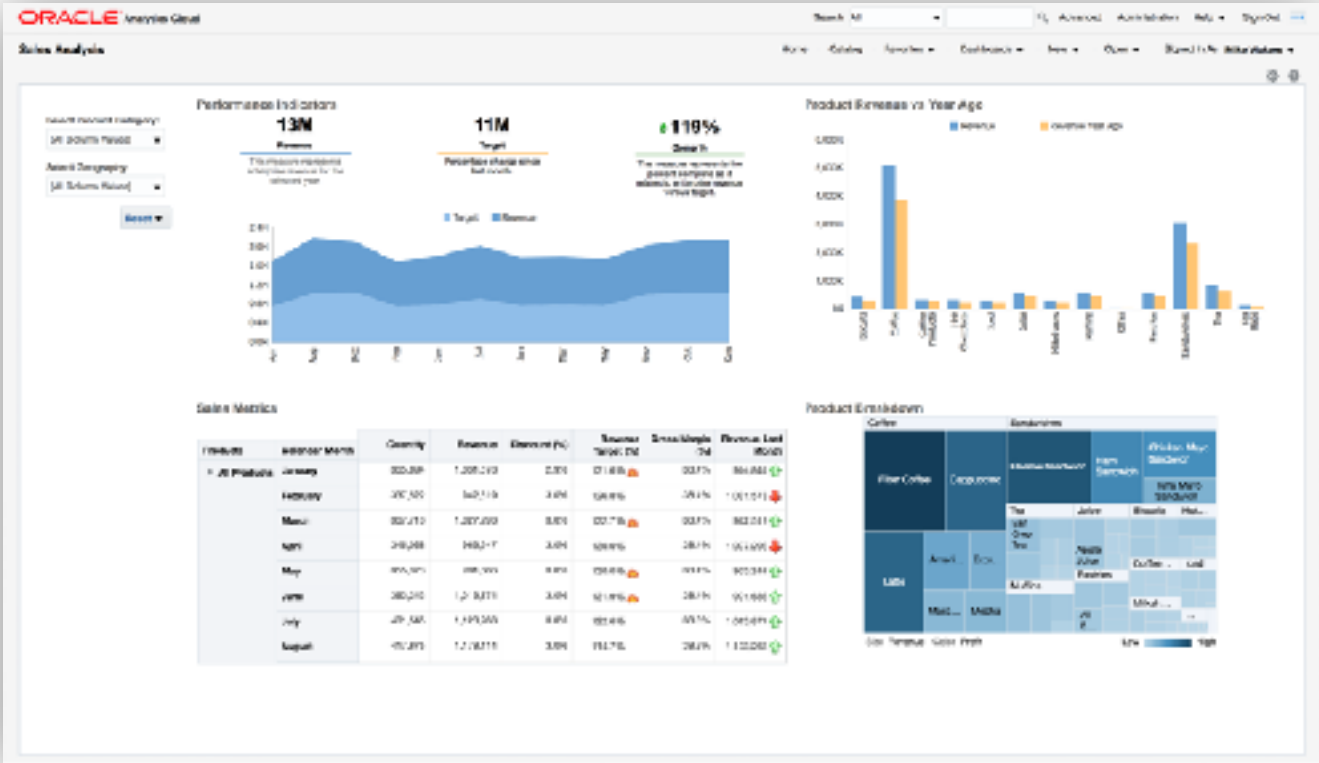


Functions

OAC supports **Every** type of analytics

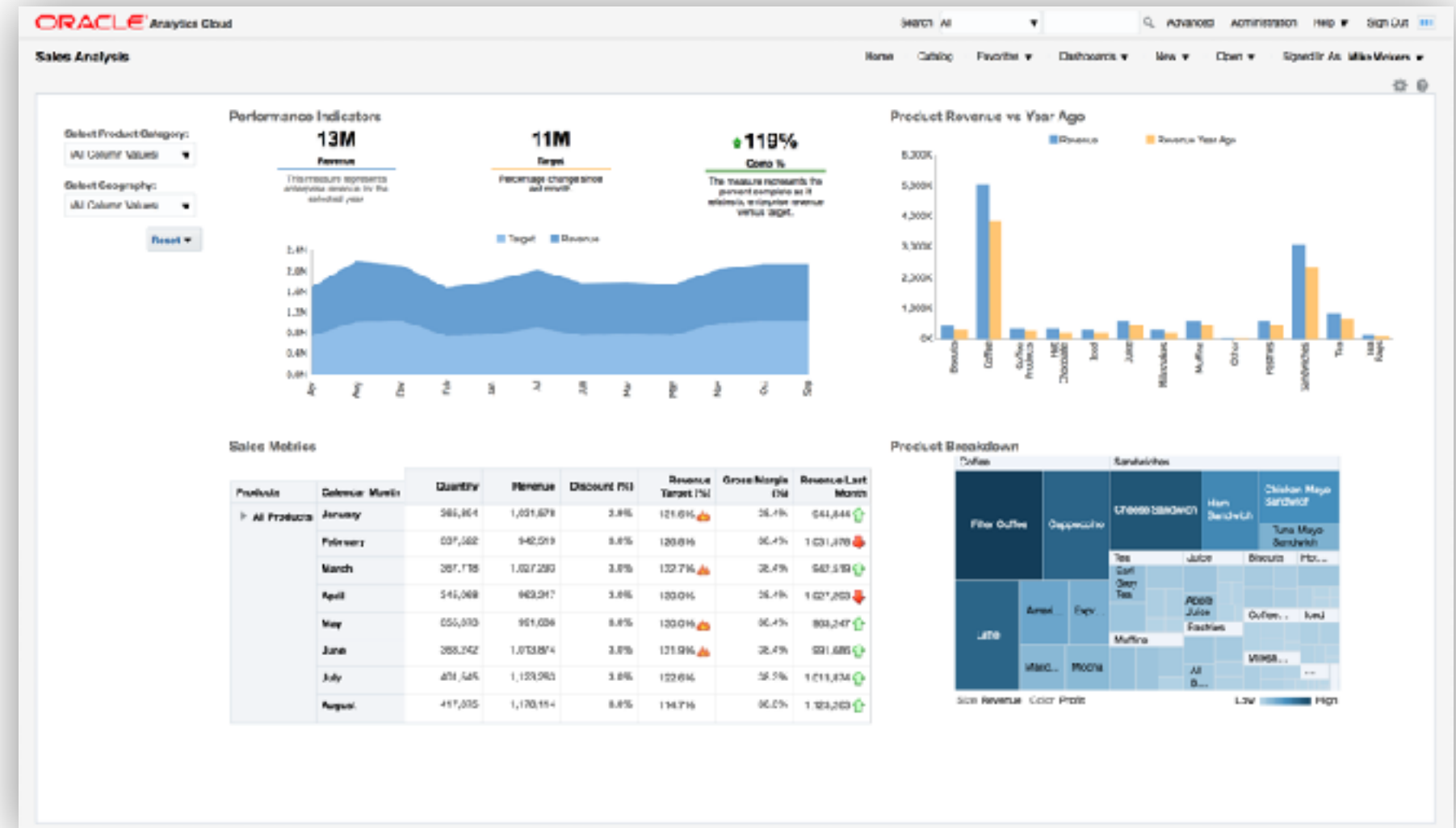
Classic

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



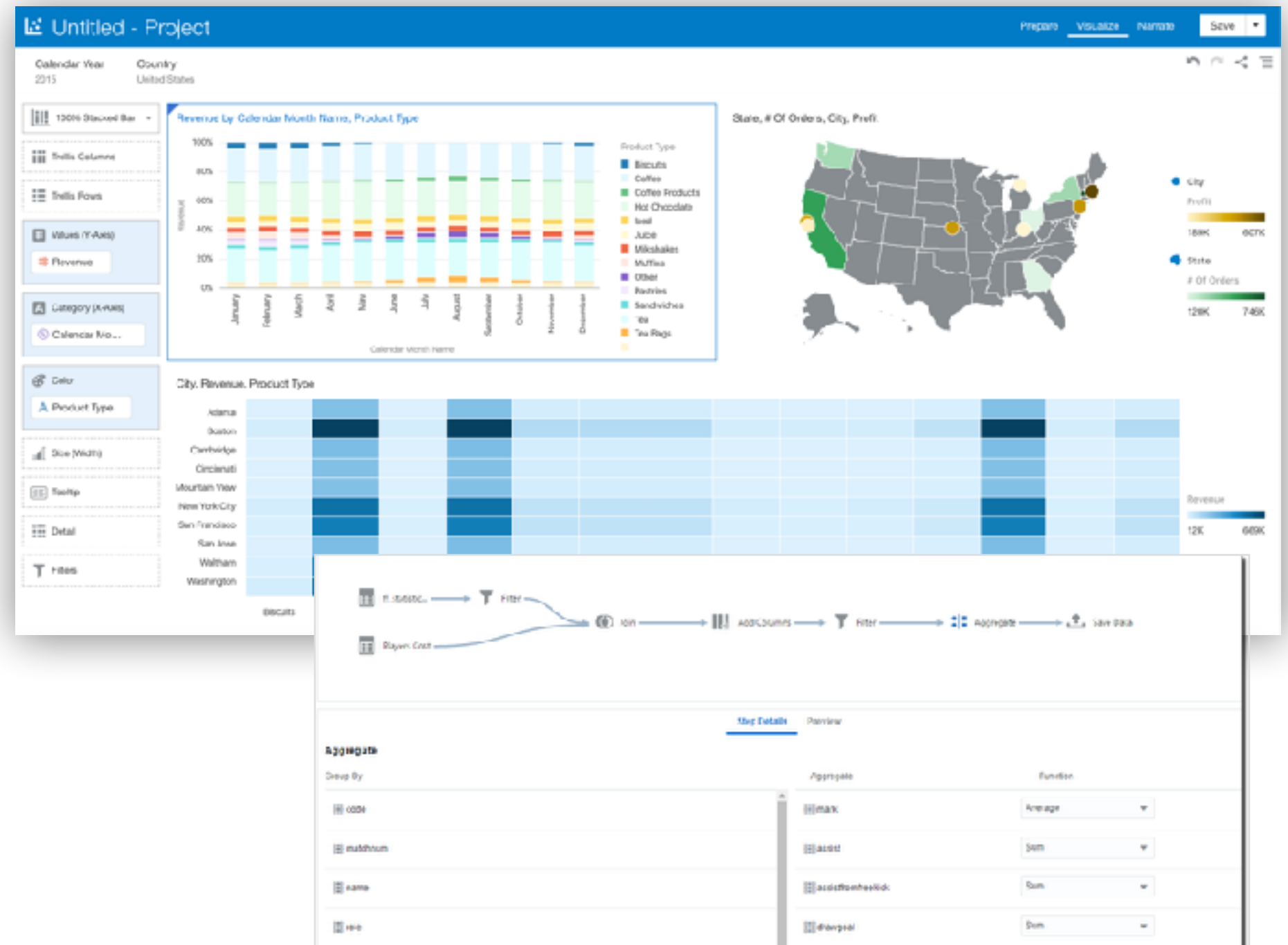
Modern Data Discovery

- 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

Unique
Source of
Truth

Data
Enrichment
and
Cleaning

Specific
Access
Control

Raw Data
To Insights

Augmented Analytics

Data Enrichment
Suggestions

Natural
Language
Processing

Explain

One-Click
Advanced
Analytics

Advanced
Machine
Learning



Data Scientist





Data Scientist

Is a person who has the knowledge and skills to conduct sophisticated and systematic **analyses of data.**

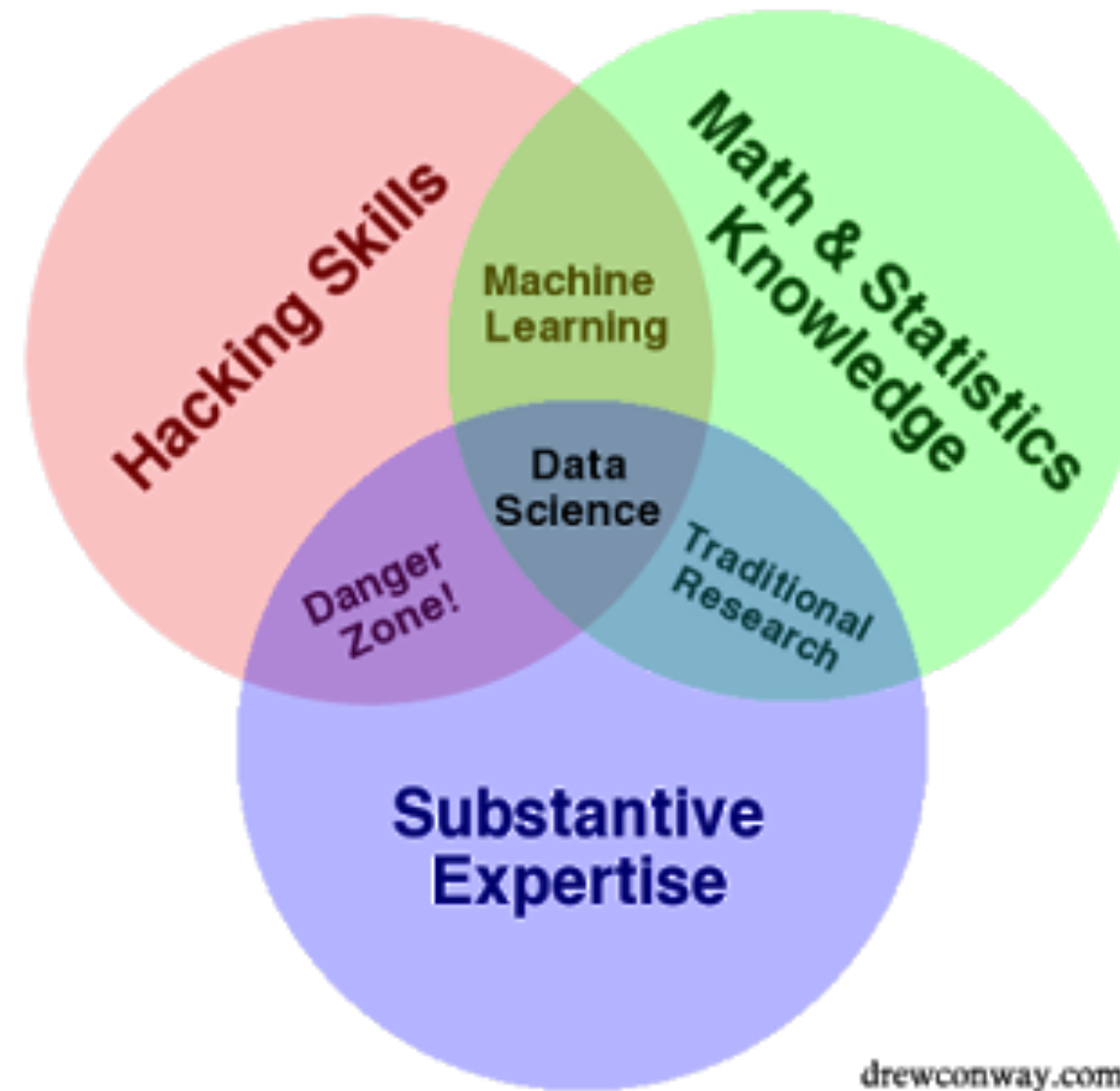
A data scientist **extracts insights** from data sets, and **evaluates and identifies strategic opportunities.**



D ata Scientist

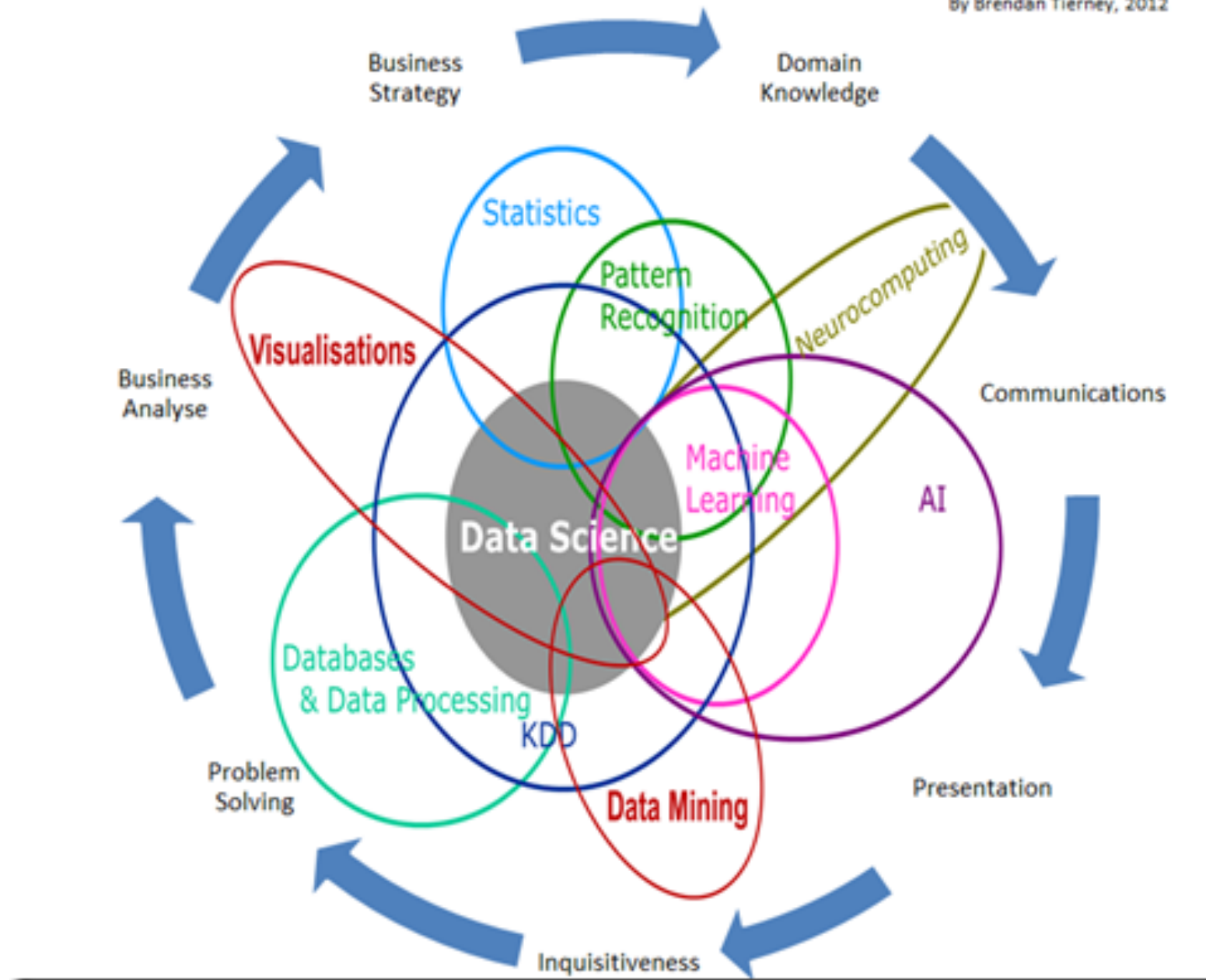
Is a Data Analyst
who lives in
California!

Data Scientist Skills



Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Brendan Tierney
Oracle Ace Director



Data Scientist



...Company Missing a Data Scientist



Low Hanging Fruit Theory

A close-up photograph of a person's hand holding a single, ripe orange fruit. The hand is positioned in the lower right quadrant of the frame, with fingers gently gripping the orange. The orange is bright orange and appears to be on a tree, as evidenced by the surrounding green leaves and branches. The background is dark and out of focus, emphasizing the hand and the fruit. The overall composition is simple and focused on the act of holding the fruit.

Democratise
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



Is this Client
going to accept
the Offer?

YES/NO

50%



Basic ML
Model

70%

Become a Data Scientist with OAC



Before Starting.... Define the Problem!





Problem Definition:
Predicting Wine Quality

TEP

T
ask



Classify
Good/Bad
Wine

E
xperience



Corpus of Wine Descriptions
with Rating

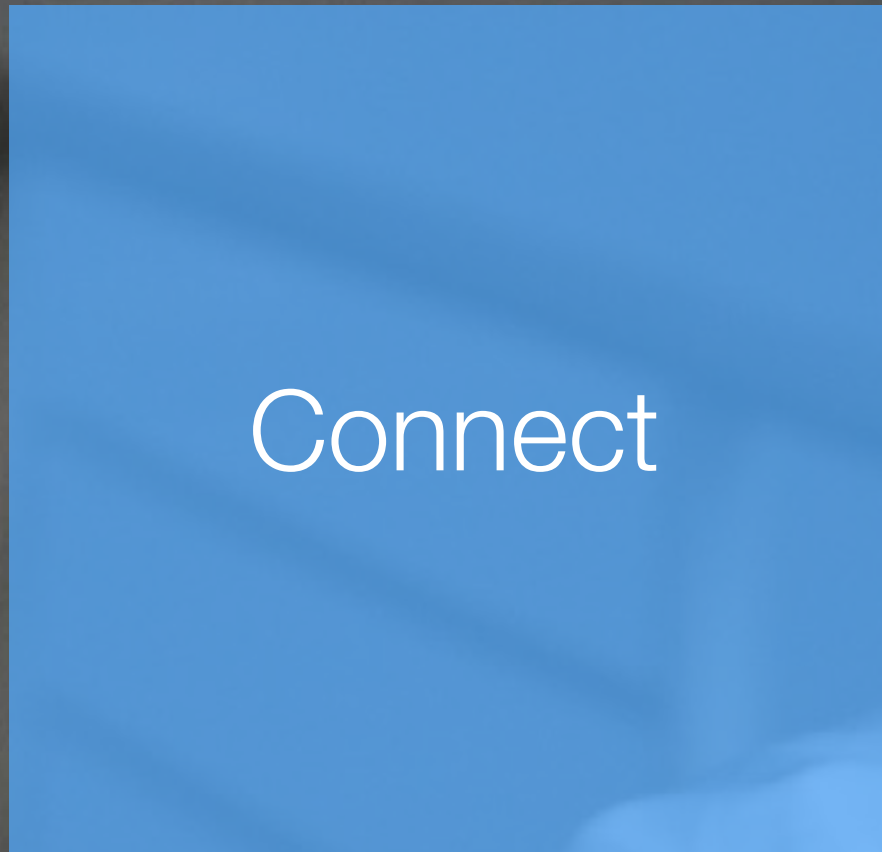
P
erformance



Accuracy

Become a Data Scientist with OAC

Connect



Select Relevant Columns and Apply Filters

The image displays three sequential screenshots of a data tool interface, illustrating the process of selecting relevant columns and applying filters.

Top Screenshot: Shows the initial state of the tool. The 'Columns' pane on the left lists available columns: **HC_ID**, **MC_EMPLOYEES**, **MC_SYSTEM**, **MC_JOB**, **MC_SUPPLIERS**, and **MC_TIME**. The **HC_ID** column is highlighted with a red box, and the **MC_TIME** column is also highlighted with a red box.

Middle Screenshot: Shows the 'Columns' pane with the **MC_TIME** column selected. The 'Columns' pane is now empty, and the 'Columns' pane on the right is populated with the selected columns: **MC_TIME**, **MC_JOB**, **MC_SUPPLIERS**, **MC_SYSTEM**, and **MC_EMPLOYEES**. The 'Columns' pane on the right is also highlighted with a red box.

Bottom Screenshot: Shows the 'Columns' pane with the **MC_TIME** column selected. The 'Columns' pane is now empty, and the 'Columns' pane on the right is populated with the selected columns: **MC_TIME**, **MC_JOB**, **MC_SUPPLIERS**, **MC_SYSTEM**, and **MC_EMPLOYEES**. The 'Columns' pane on the right is also highlighted with a red box.

The interface includes a 'Columns' pane on the left, a 'Columns' pane on the right, and a 'Columns' pane at the bottom. The 'Columns' pane on the left lists available columns, and the 'Columns' pane on the right lists selected columns. The 'Columns' pane at the bottom lists the columns selected for the current view.

A_EMP_ID	D_EMP	A_JOB	A_JOB_ID	A_JOB_NAME	A_JOB_LVL	A_JOB_GRP	A_YEAR
10001	10001	10001	10001	10001	10001	10001	10001
10002	10002	10002	10002	10002	10002	10002	10002
10003	10003	10003	10003	10003	10003	10003	10003
10004	10004	10004	10004	10004	10004	10004	10004
10005	10005	10005	10005	10005	10005	10005	10005
10006	10006	10006	10006	10006	10006	10006	10006
10007	10007	10007	10007	10007	10007	10007	10007
10008	10008	10008	10008	10008	10008	10008	10008
10009	10009	10009	10009	10009	10009	10009	10009
10010	10010	10010	10010	10010	10010	10010	10010
10011	10011	10011	10011	10011	10011	10011	10011
10012	10012	10012	10012	10012	10012	10012	10012
10013	10013	10013	10013	10013	10013	10013	10013
10014	10014	10014	10014	10014	10014	10014	10014
10015	10015	10015	10015	10015	10015	10015	10015
10016	10016	10016	10016	10016	10016	10016	10016
10017	10017	10017	10017	10017	10017	10017	10017
10018	10018	10018	10018	10018	10018	10018	10018
10019	10019	10019	10019	10019	10019	10019	10019
10020	10020	10020	10020	10020	10020	10020	10020

PROJECT	DEPARTMENT
10001	10001
10002	10002
10003	10003
10004	10004
10005	10005
10006	10006
10007	10007
10008	10008
10009	10009
10010	10010
10011	10011
10012	10012
10013	10013
10014	10014
10015	10015
10016	10016
10017	10017
10018	10018
10019	10019
10020	10020

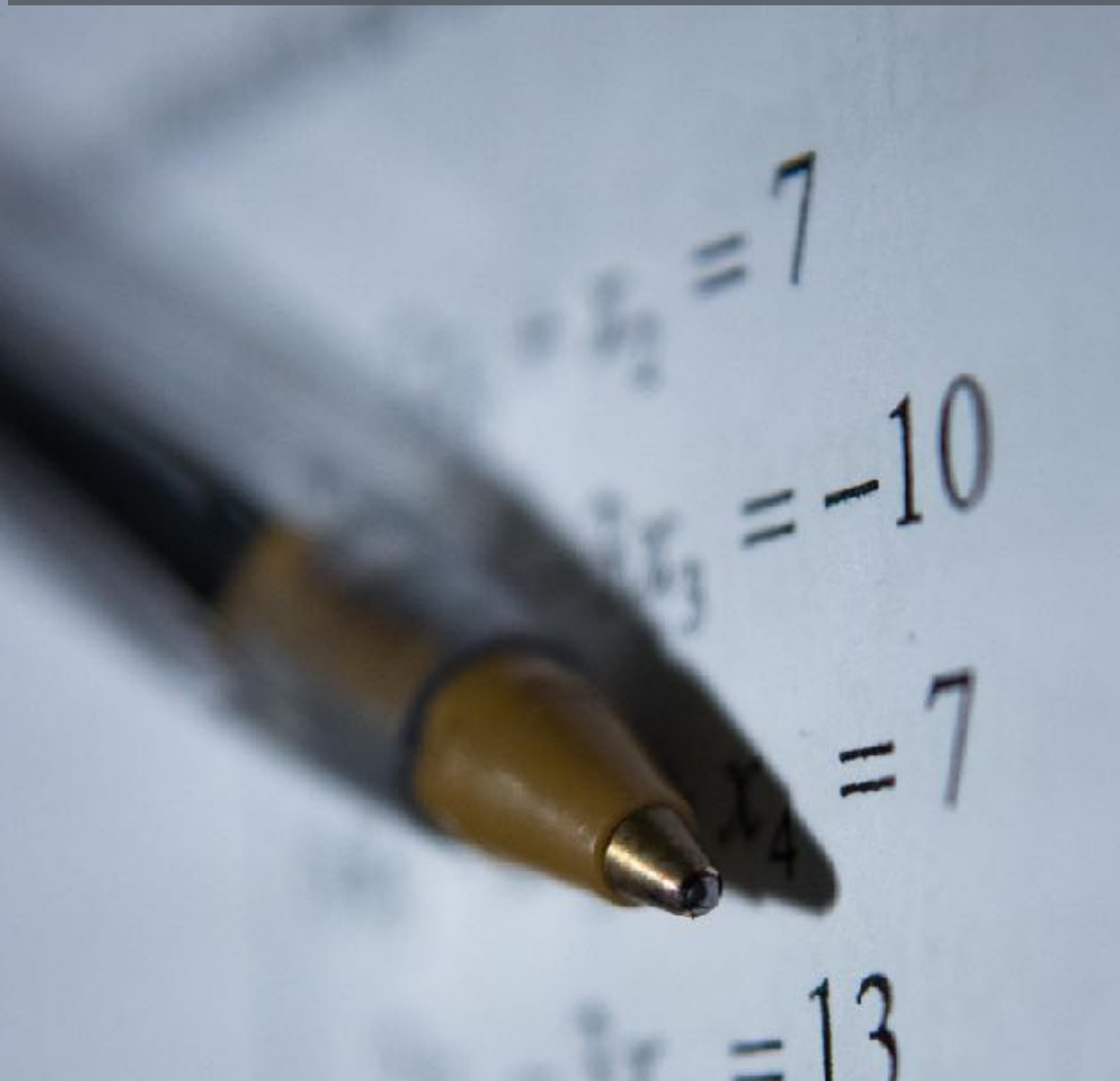
Become a Data Scientist with OAC

Connect

Clean



What Everybody Thinks
a Data Scientist Does



What He Really Does



 IDG CONTRIBUTOR NETWORK [Want to Join?](#)

THE COGNITIVE CODER

By **Armand Ruiz**, Contributor, InfoWorld | SEP 26, 2017

Opinions expressed by ICN authors are their own.

The 80/20 data science dilemma

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, which is an inefficient data strategy

Cleaning What?

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

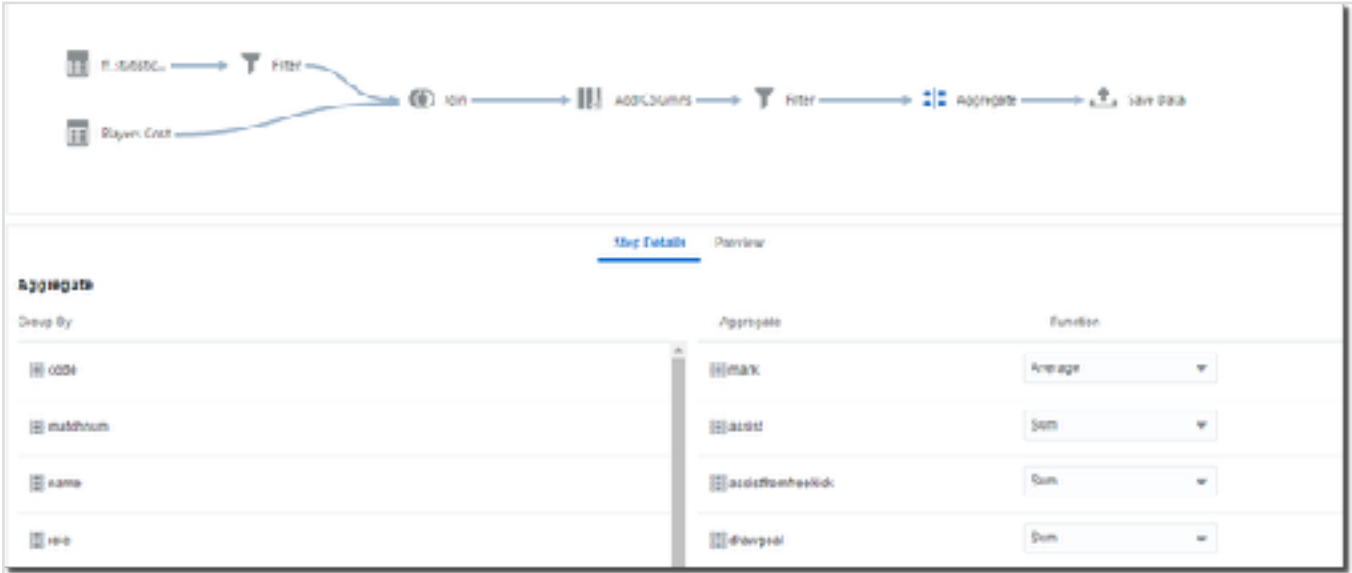
Train: 80%
Test: 20%

Train/Test Set Split

Cleaning How?

Data Flows

- Filter
- Aggregate
- Join



A grid of data processing tool icons. The top row includes: Add Data, Join, Merge Rows, Filter, Aggregate, Save Data, and Create Essbase Cube. The second row includes: Add Columns, Select Columns, Rename Columns, Merge Columns, Bin, Group, and Branch. The third row includes: Cumulative Value, Time Series Forecast, Analyze Sentiment, and Apply Custom Script. The bottom row includes: Train Numerical Prediction, Train Multi-Classifer, Train Binary Classifier, Train Clustering, Train Custom Model, and Apply Model.

Cleaning What?

CASE ...
Automated
WHEN...

Missing Values

UPPER

Mark <> MArk

Wrong Values

City
FILTER
"Rome"

Irrelevant Observations

COLUMN
RENAME

Col1 > Name

Labelling Columns

FILTER?

Role: CIO
Salary: 500 K\$

Handling Outliers

KPI/
Automated
(MAX-MIN)

Feature Scaling

of Clicks
COUNT

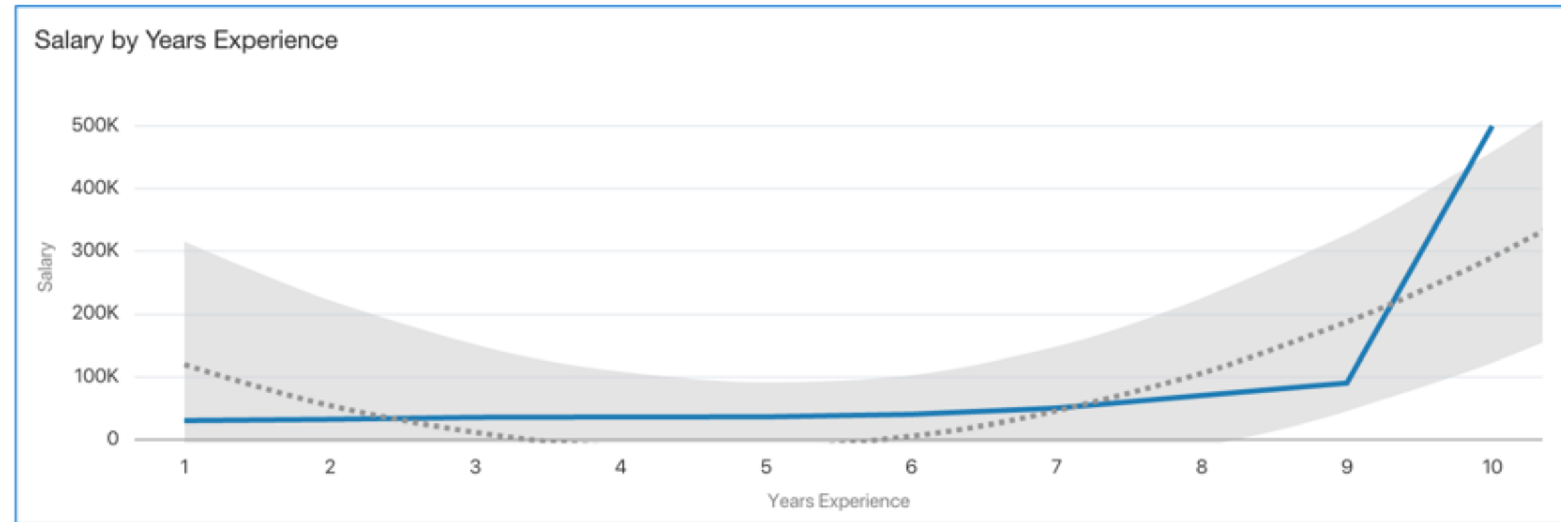
Aggregation

FILTER
Automated

Train/Test Set Split

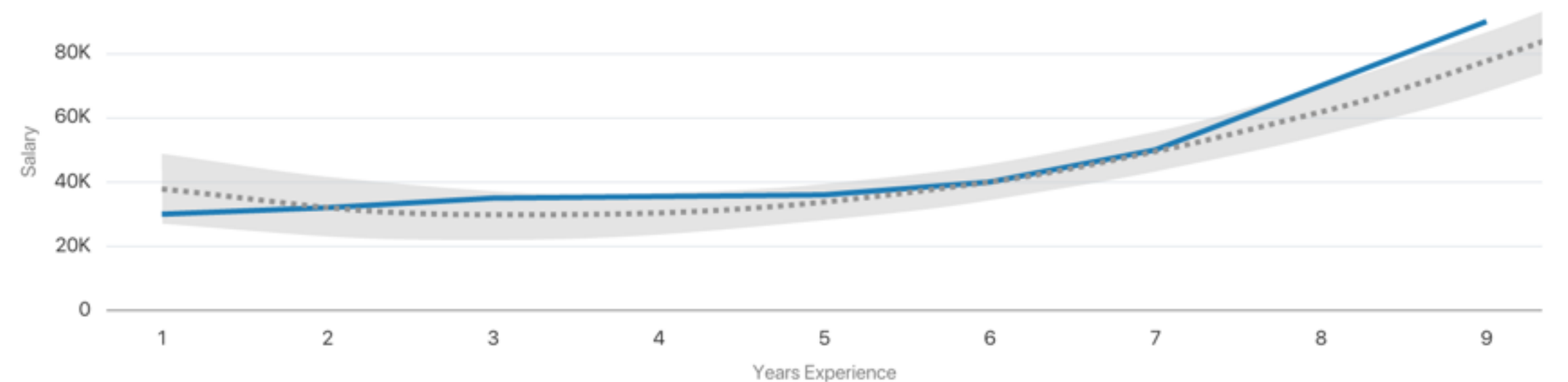
Why Removing an Outlier?

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



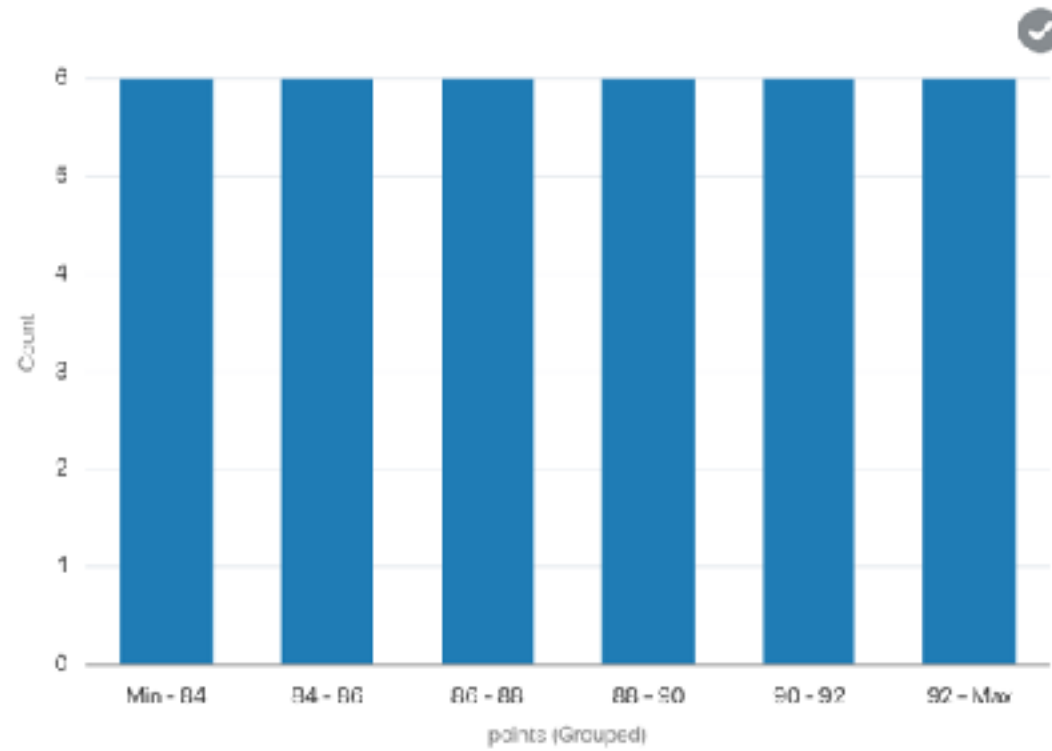
Salary by Years Experience

Years-Experience: 10

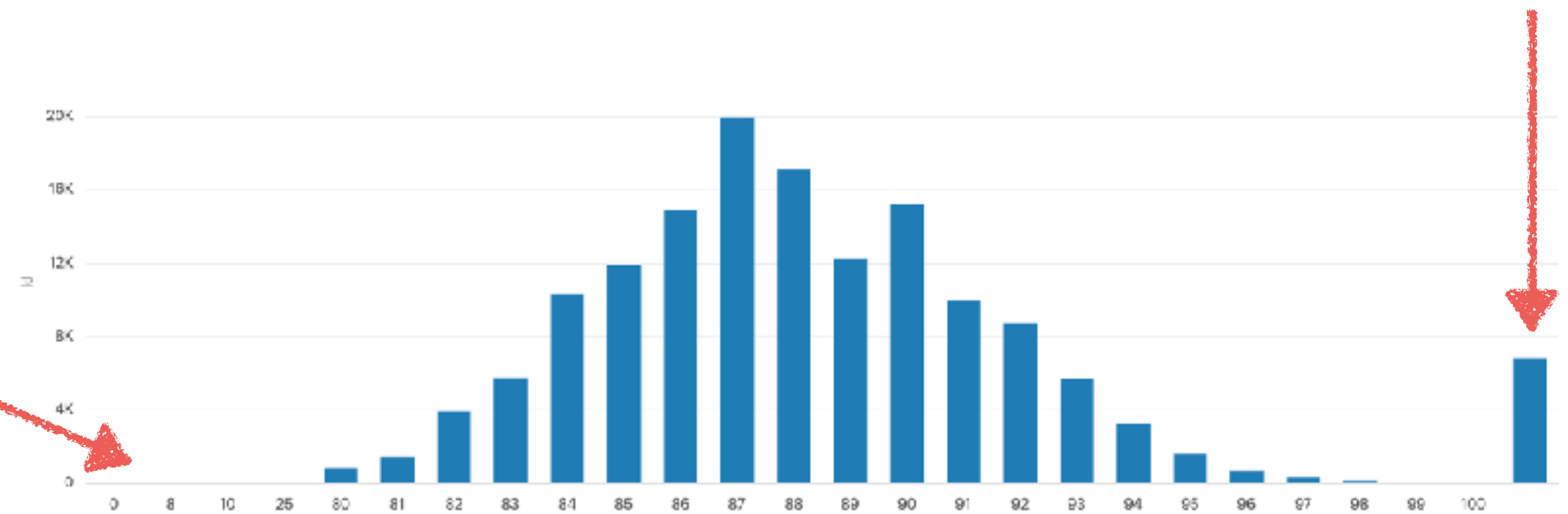


How To Find Outliers? One Dimension

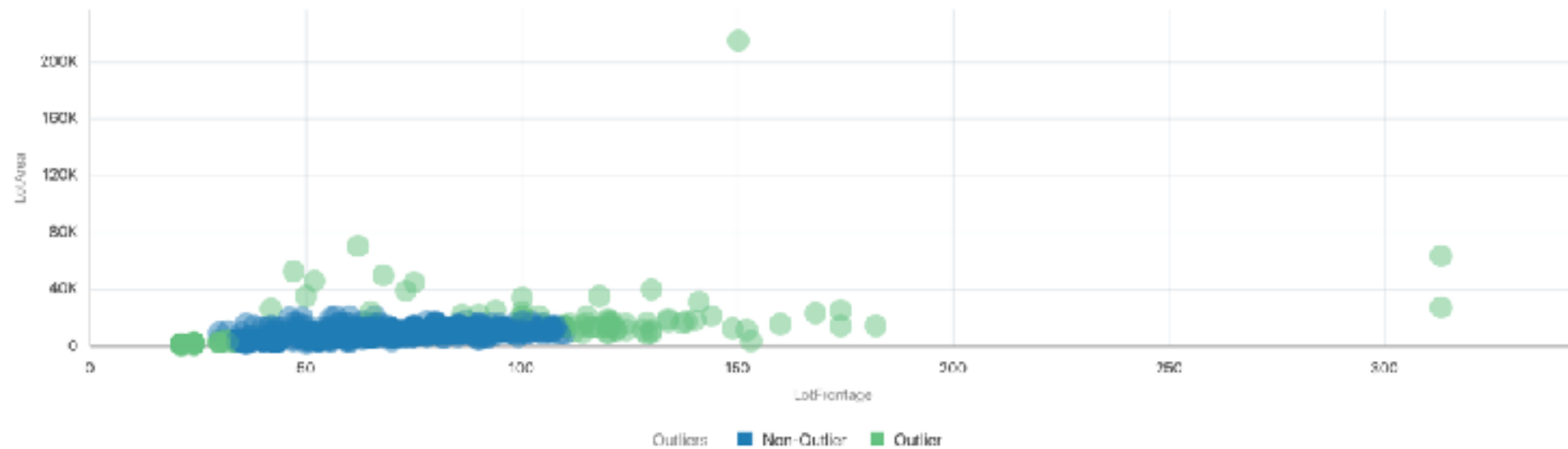
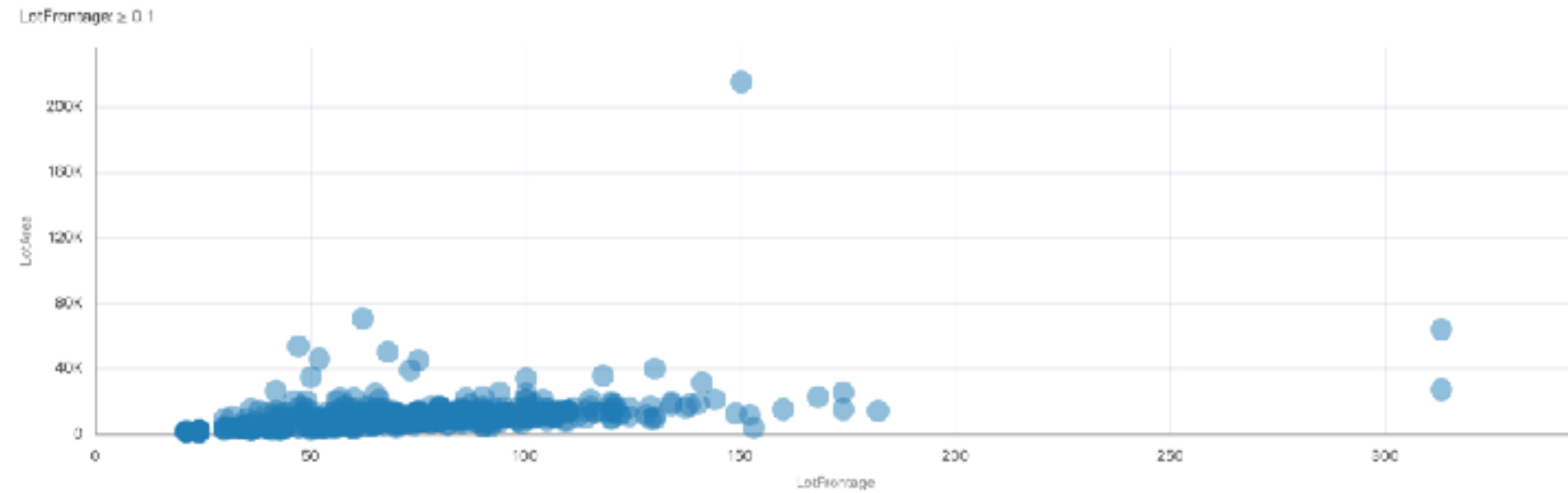
Basic facts about points



points is a Numeric Measure, whose sum across 130,696 rows is 11,471,024.00. The values of points on each row range from 0.00 to 100.00 and is 87.00 on average.



How To Find Outliers? Two Dimensions



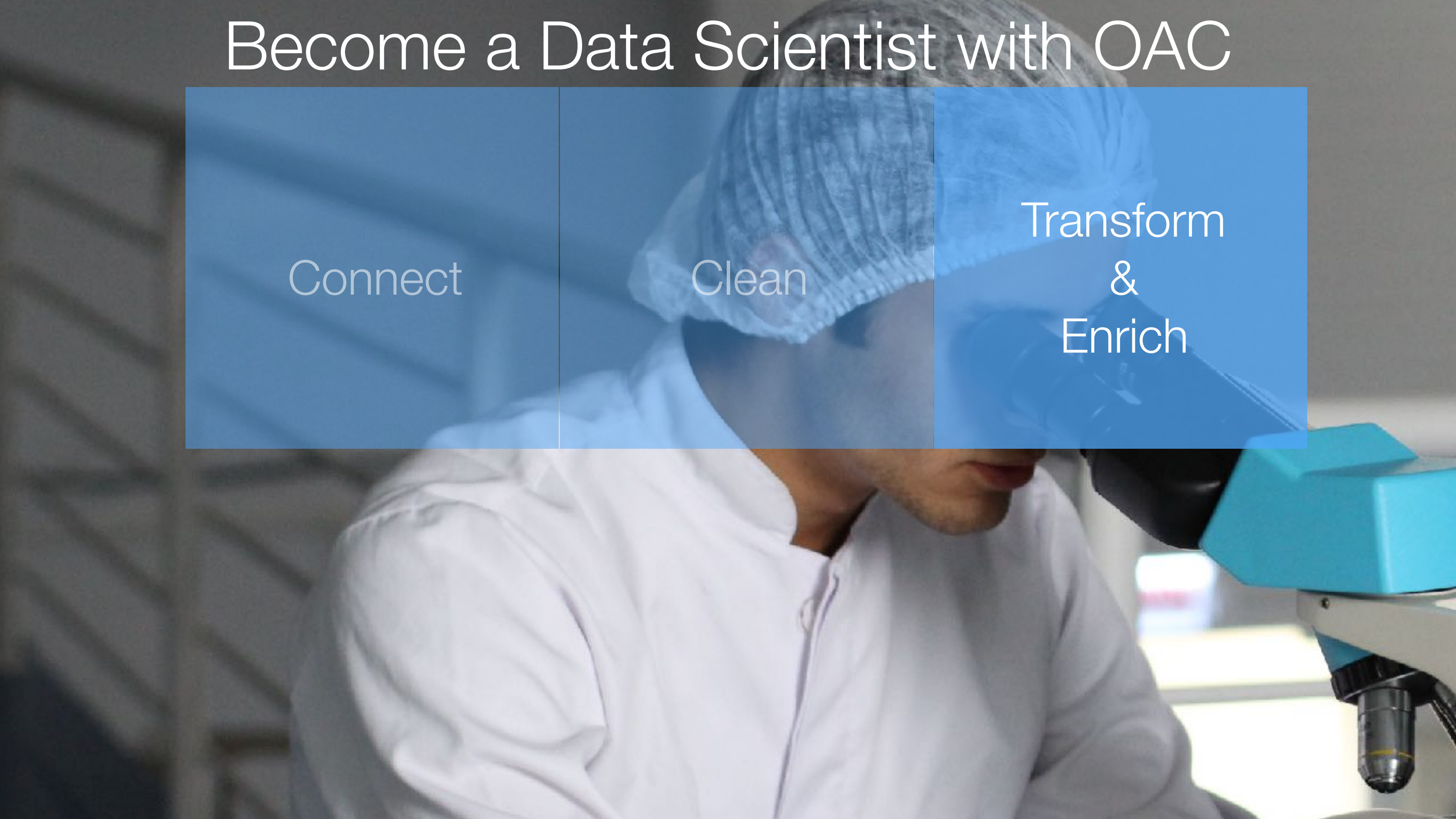
- Undo (Ctrl+Z)
- Add Clusters
- Add Outliers**
- Add Reference Line
- Add Trend Line
- Use as Filter
- Duplicate Visual
- Copy Visual
- Copy Data to Clipboard
- Delete Visual
- Canvas Layout ▶

Become a Data Scientist with OAC

Connect

Clean

Transform
&
Enrich



Feature Engineering

Location -> ZIP Code

Additional
Data Sources?

Name -> Sex

2 Locations -> Distance

Data Flow

Day/Month/Year -> Date

Data Preparation Recommendations

The screenshot displays a data preparation interface with a central table of wine data. The table has columns: Id, country, country_country_name, country_iso3, and description. The 'country' column is highlighted in blue, and the 'country_country_name' column is highlighted in orange. To the left, a 'Preparation Script' panel shows a list of enrichment actions, with 'Enrich country with ...' selected. To the right, a dropdown menu for 'country (16)' lists various enrichment recommendations, such as 'Enrich country with iso2', 'Enrich country with iso_numeric', 'Enrich country with fips', 'Enrich country with capital', 'Enrich country with square_km', 'Enrich country with population', 'Enrich country with continent', 'Enrich country with tld', 'Enrich country with currency_abbrev', and 'Enrich country with currency_name'.

Id	country	country_country_name	country_iso3	description
1020	Germany	Germany	DEU	Pretty floral
4971	France	France	FRA	The chalk st
1511	Austria	Austria	AUT	Ivy leaf and
4055	US	United States	USA	You'll taste t
3641	US	United States	USA	Intense arcn
3852	US	United States	USA	From young
6	Spain	Spain	ESP	Slightly gritt
1013	France	France	FRA	A dry wine v
1122	US	United States	USA	Winemakers
9	US	United States	USA	The produce
2368	US	United States	USA	This densely
11	US	United States	USA	From 18-yea
4838	Portugal	Portugal	PRT	This is a pov
3814	Greece	Greece	GRC	Grapefruit, le

Spatial Enrichment

The screenshot displays the Oracle Spatial Studio interface. The top panel is titled "Spatial Analysis Operations" and includes tabs for "All", "Filter", "Combine", "Transform", and "Measure". A search bar is located in the top right corner. Below the tabs, five operation cards are visible:

- Add a buffer of a specified distance**: SDO_GEOM.SDO_BUFFER. No information.
- Create point in the middle of a shape**: SDO_GEOM.SDO_CENTROID. More information.
- Create the area defined by the "rubber band" envelope of a shape**: SDO_GEOM.SDO_CONVEHULL. More information.
- Create the smallest circle that encloses a shape**: SDO_GEOM.SDO_MBC. More information.
- Create the smallest box that encloses a shape**: SDO_GEOM.SDO_MBR. More information.

The bottom panel shows a map of the United States of America with a legend overlay. The legend includes the following items:

- TEST_CUSTOMERS (blue dot)
- TEST_SHOPS (red dot)
- TEST_SHOPS BUFFER (grey square)
- TEST_CUSTOMERS WIT... (green dot)

The map shows blue dots representing customers, red dots representing shops, and grey squares representing the buffer around the shops. Green dots represent customers within a certain distance of the shops. The text "UNITED STATES OF AMERICA" is visible on the map. The bottom right corner of the map area contains the text "© OpenMapTiles © OpenStreetMap contributors".

Oracle Spatial Studio

<https://www.rittmanmead.com/blog/2019/07/oracle-spatial-studio/>

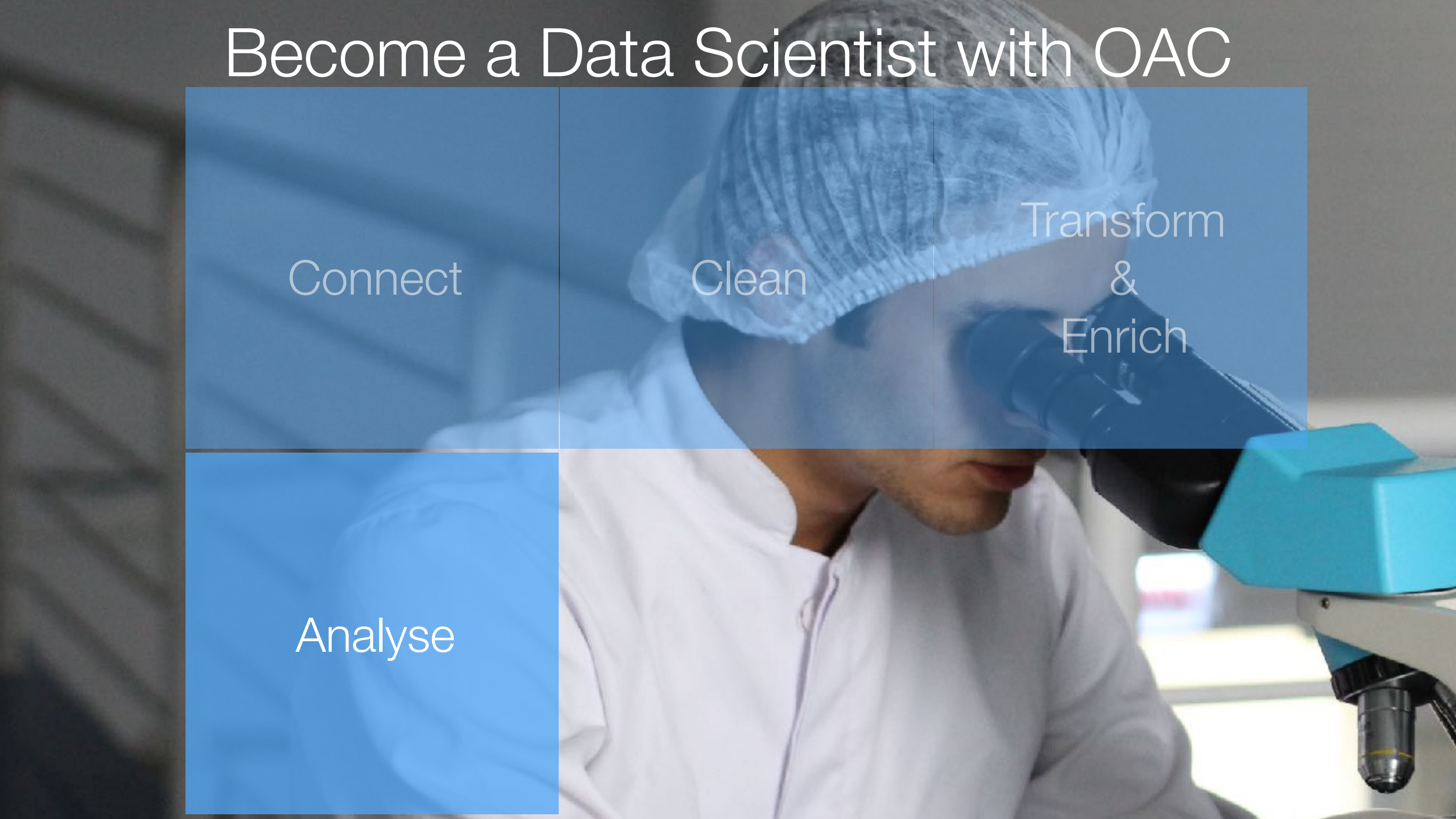
Become a Data Scientist with OAC

Connect

Clean

Transform
&
Enrich

Analyse



Data Overview

Results

Metadata



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

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 this data best explain the values of points?

points is a Numeric Measure, whose average across 150,935 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.

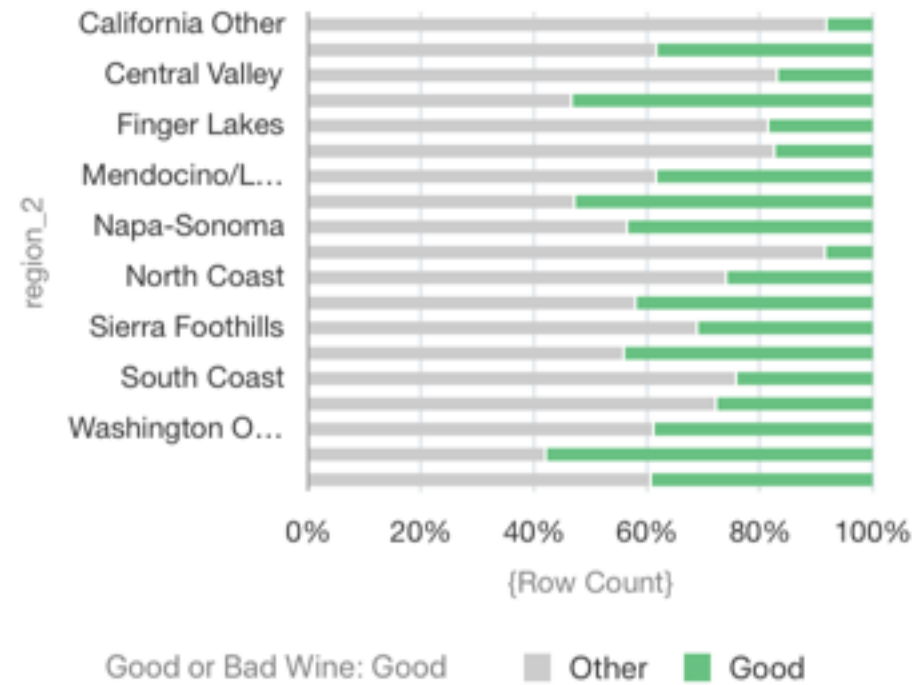
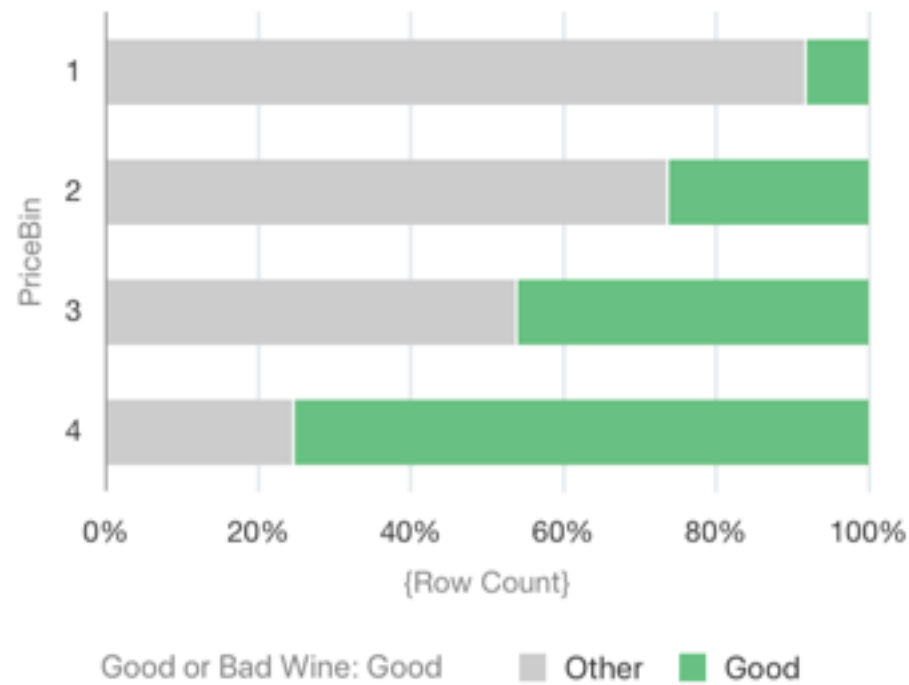
- id
- country
- country_continent
- country_latitude
- country_longitude
- country_timezone

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: **PriceBin, region_2**

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



Become a Data Scientist with OAC

A scientist in a white lab coat and blue hairnet is looking through a blue and black microscope. The background is a blurred laboratory setting. The image is overlaid with a blue grid that divides it into six sections, each containing a step of the OAC process.

Connect

Clean

Transform
&
Enrich

Analyse

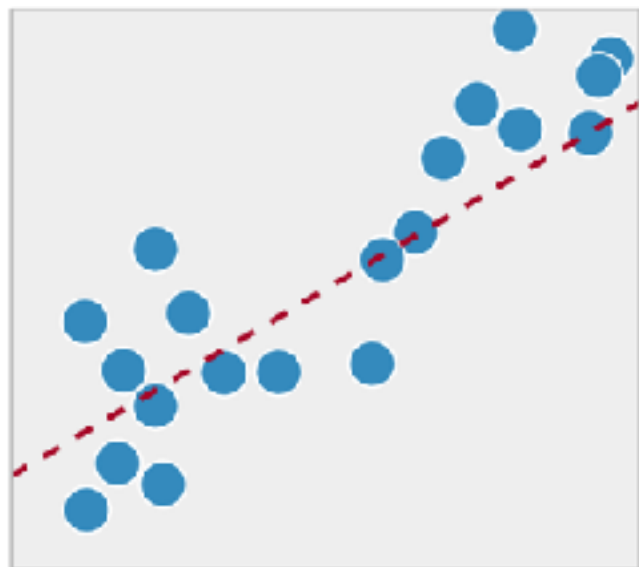
Train
&
Evaluate

What Problem are we Trying to Solve?

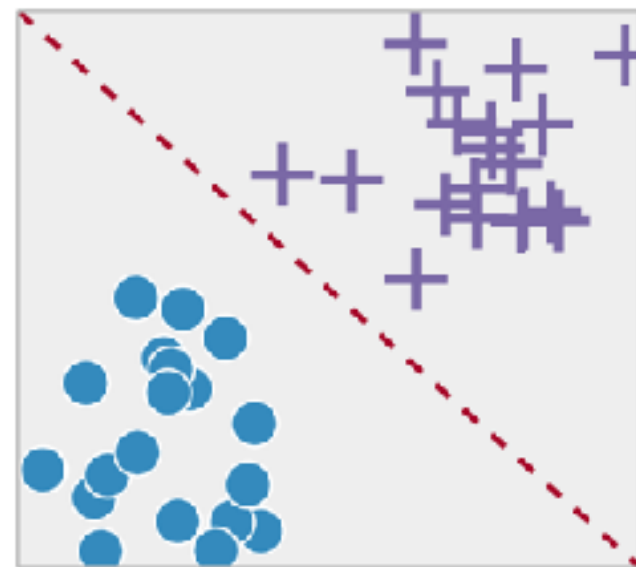
Supervised

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

Regression



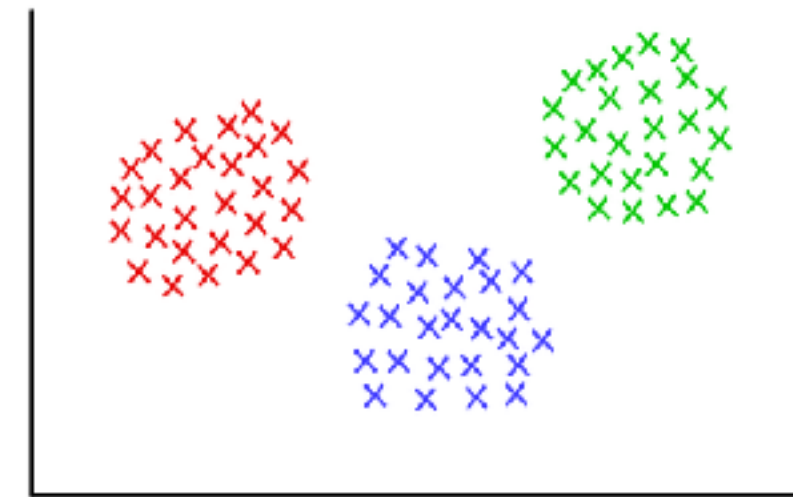
Classification



Unsupervised

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

Clustering



Easy Models

Scatter

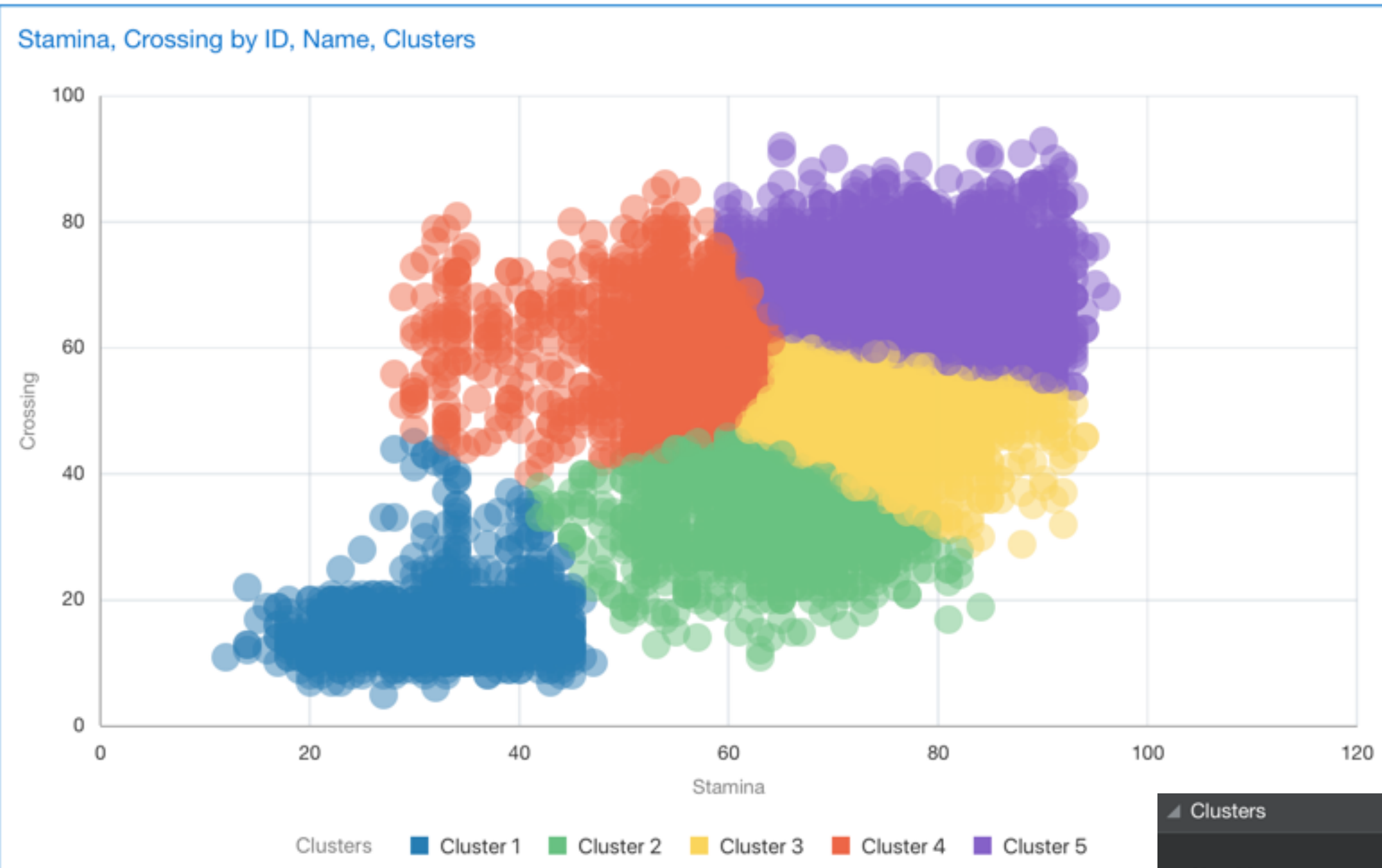
Trellis Columns

Trellis Rows

Values (Y-Axis)
Crossing

Values (X-Axis)
Stamina

Category (Points)
ID
Name



Clusters

Algorithm K-Means

Groups

NLP

Language Narrative

Attributes

Name

Values

Stamina

Crossing

Filters

Stamina, Crossing by Name

The data compares the Stamina with the Crossing for a total of 997 Names.

Focus on Stamina

- When taken together, the 997 Names reach a total value of 66,234, an average of 66.43. The most frequent value is 68 and appears 35 times.
- The data was able to be divided into 4 distinct categories.
 - A. González is the biggest, with a Stamina of 313.
 - A. Ba and A. Correa are the next two in terms of Stamina, with 259 on average (0.78% of the total Stamina, about 0.39% each).
 - A. Al Khaibari, A. Majrashi and A. Castro are the next three in terms of Stamina, with 206.67 on average (0.94% of the total Stamina, about 0.31% each).
 - A. Mosquera, A. Diallo, A. Davies and 988 others finish the list, with 65.37 on average. This last group makes up the majority of Names (97.81% of the total Stamina, about 0.1% each).

Focus on Crossing




- When taken together, the 997 Names amount to a total value of 51,410, 51.57 on average. The most frequent value is 65 and appears 37 times.
- The data was able to be divided into 4 distinct categories.
 - A. González is the largest, with a Crossing of 263.
 - A. Castro is the second biggest, with a Crossing of 202.
 - A. Correa's numbers were not as high, but it is the third most important, with a Crossing of 172.
 - The remaining Names, A. Ba, A. Gómez, A. Majrashi and 991 others, finish the list, with 51.08 on average. Combined, this last group contains the majority of Names (98.76% of the total Crossing, approximately 0.1% each).





The comparison of two unordered measures is not yet available. Unordered means that the data is not in chronological order. The application will generate a separate analysis for each measure. Stay tuned, future releases will add functionalities for unordered dimensions.

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?


Select Train Numeric Prediction Model Script



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

Regression Method **Lasso**

Method for linear regression training.

Regularization Weight **1**

Regularization Weight(L1 Ratio or L2 Ratio). Please enter 0 if it is Ordinary Least Squares linear regression.

Categorical Column Imputation **Most Frequent**

The mode method for categorical features to fill NA. Two options: most frequent and least frequent. Default is most frequent.

Numerical Column Imputation **Mean**

The mode method for numeric features to fill NA. Four options: mean, max, min, median. Default is mean.

Categorical Encoding Method **Indexer**

Encoding method.

Maximum Null Value Percent **80**

Maximum Null Value Percent

Train Partition Percent **80**

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: 0.6
L2 Ratio

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

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

Data Sets | Connections | **Data Flows** | Sequences

Type	Name
>>>	ELN1
>>>	LR2
>>>	LR1

Machine Learning

Scripts | **Models**

Type	Name
	ELN1
	LR2
	LR1

Compare - Classification

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

		Predicted Value	
		Good	Bad
Real Value	Good		
	Bad		

There is No Single Truth...

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

$$502 / (502 + 896) = 64.09\%$$

Precision

$$471 / (471 + 866) = 64.77\%$$

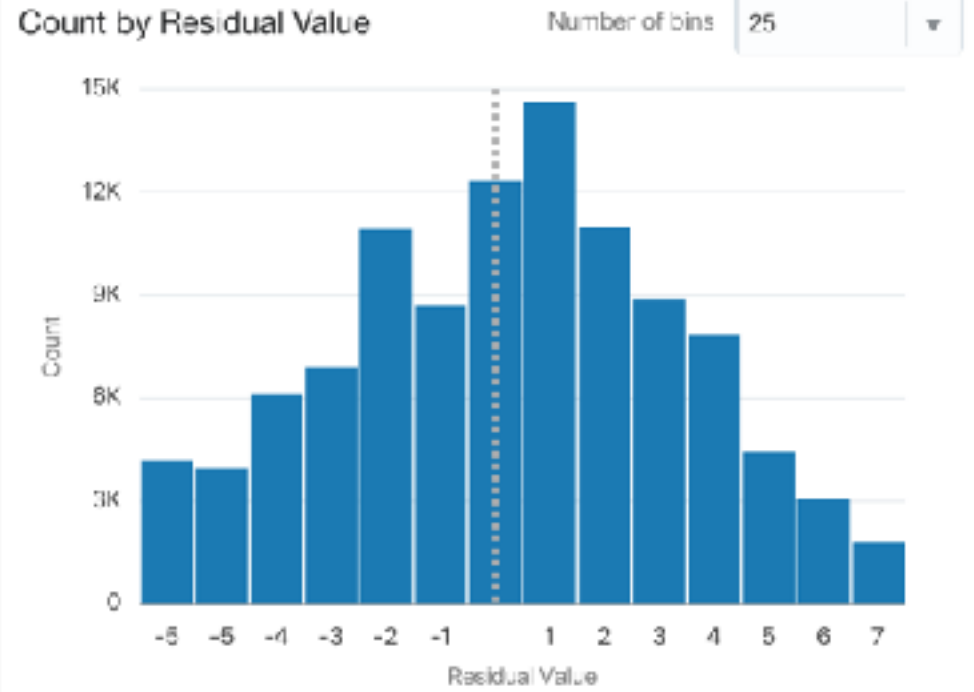
		Predicted Values		Total
		0.0	1.0	
Actual Values	0.0	40439	471	40910 (90%)
	1.0	3761	866	4627 (10%)
	Total	44200 (97%)	1337 (3%)	45537 (100%)

Compare - Regression

LR1
Numeric Prediction Model

Save Close

- General
- Quality
- Related

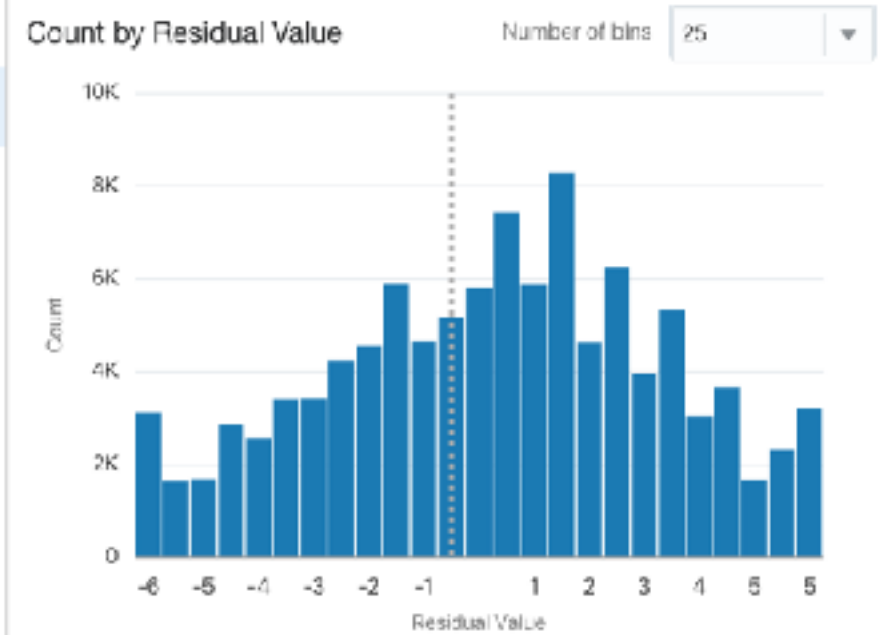


Mean Absolute Error (MAE) 2.60
Median Absolute Error 2.24
Root Mean Squared Error 3.22
Relative Absolute Error (RAE) 1.00
Relative Squared Error (RSE) 1.00
Coefficient of Determination (R^2) 0%

LR2
Numeric Prediction Model

Save Close

- General
- Quality
- Related



Mean Absolute Error (MAE) 2.58
Median Absolute Error 2.32
Root Mean Squared Error 3.16
Relative Absolute Error (RAE) 0.98
Relative Squared Error (RSE) 0.98
Coefficient of Determination (R^2) 4%

Become a Data Scientist with OAC

Connect

Clean

Transform
&
Enrich

Analyse

Train
&
Evaluate

Predict

Use On the Fly

+
Add Data Set...
Create Scenario...
Add Custom...

Create Scenario - Select Model

Search

Type	Name
	BinaryCart2
	BinaryCart1
	BinaryLogistic1
	ELN1
	LR2
	LR1

Edit Scenario - Map Your Data

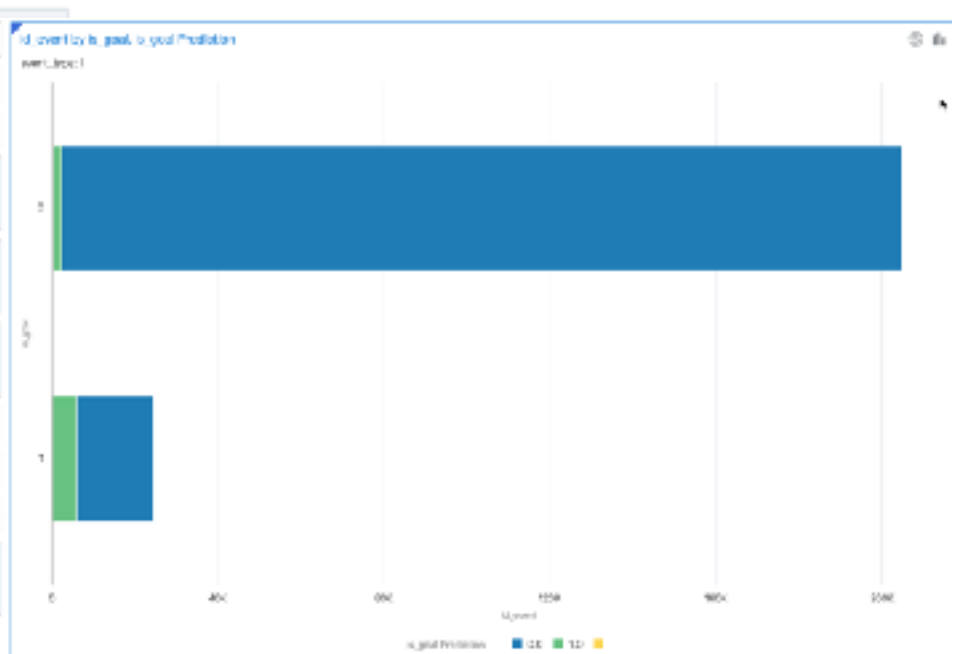
Select which Data Set you want to use with the Model

Data Set:

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

Model Input	Map To
bodypart	bodypart
location	location
player	player
situation	situation
is_goal	is_goal

* Required Fields



Step of a Data Flow



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
<input checked="" type="checkbox"/>	PredictedValue	<input type="text" value="PredictedValue"/>
<input checked="" type="checkbox"/>	PredictionConfidencePercentage	<input type="text" value="PredictionConfidencePercentage"/>
<input checked="" type="checkbox"/>	PredictionGroup	<input type="text" value="PredictionGroup"/>

Inputs

Model	Input
situation	<input type="text" value="situation"/>
bodypart	<input type="text" value="bodypart"/>
player	<input type="text" value="player"/>

Congratulations!

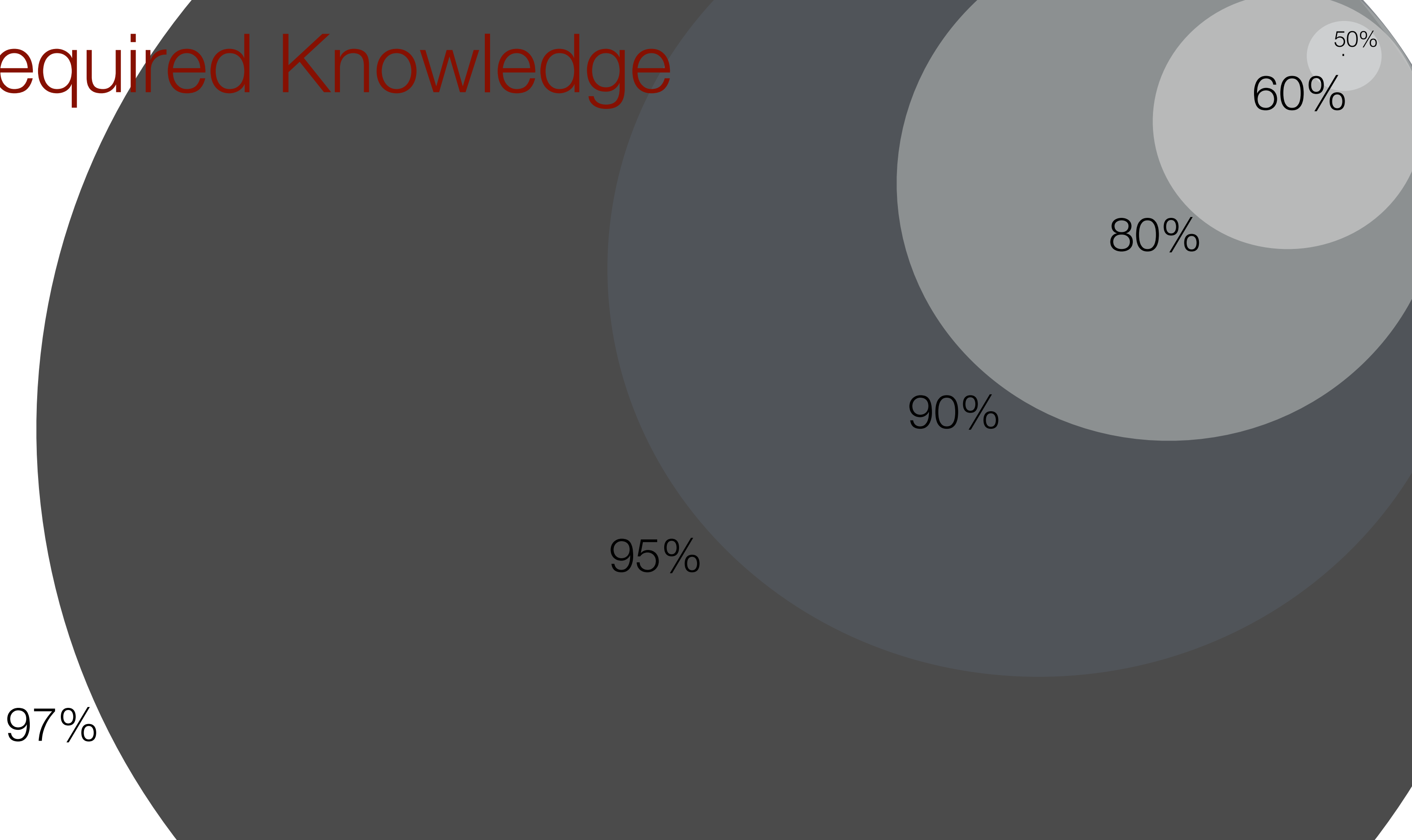


...You are now a Data Scientist!



Nearly
There

Required Knowledge



...But

Data Cleaning



Feature
Engineering

Model Creation &
Evaluation

Feature
Selection

80% > 50%



ML Production Deployment

Data Scientist

ML -> Data

Oracle Advanced Analytics



Become a Data Scientist with OAC

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

ML in Action with OAC

A chalkboard sign is the central focus, mounted on a white pillar. The sign is dark green and has the text 'Hello Wine, Goodbye Problems' written in white chalk. The background shows an outdoor restaurant patio at night, with tables, chairs, and warm lighting. The patio is decorated with string lights and hanging lanterns. The overall atmosphere is cozy and inviting.

Hello
Wine,
Goodbye
Problems

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



Insights Lab

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

Data Science



OAC

Become a Data Scientist

Francesco Tisiot
Analytics Tech Lead

rittmanmead
A DATA AND ANALYTICS COMPANY

