The other stuff: Turning machine learning into data science

Brandon Rohrer
Senior Data Scientist

Microsoft

The Secret to Data Science

Know your data.
Know your question.
Know your domain.
Know how you're going to use the answer.

Machine Learning answers a question

Prediction: What value?

Classification: What class?

Clustering: Which group?

Anomaly detection: Is it weird?

Recommendation: Which option?

Machine Learning alone is not Data Science

1. Data quality: Is it garbage?
2. Data exploration: What does it look like?
3. Feature Engineering: What does it mean?
4. Operationalization: How do I harvest data and deliver answers?

Data exploration

Distribution

Structure

Most relevant features

Visualization

Know your data.

Mystery data

Two features

Four classes

1600 data points

rows 1600 columns 3

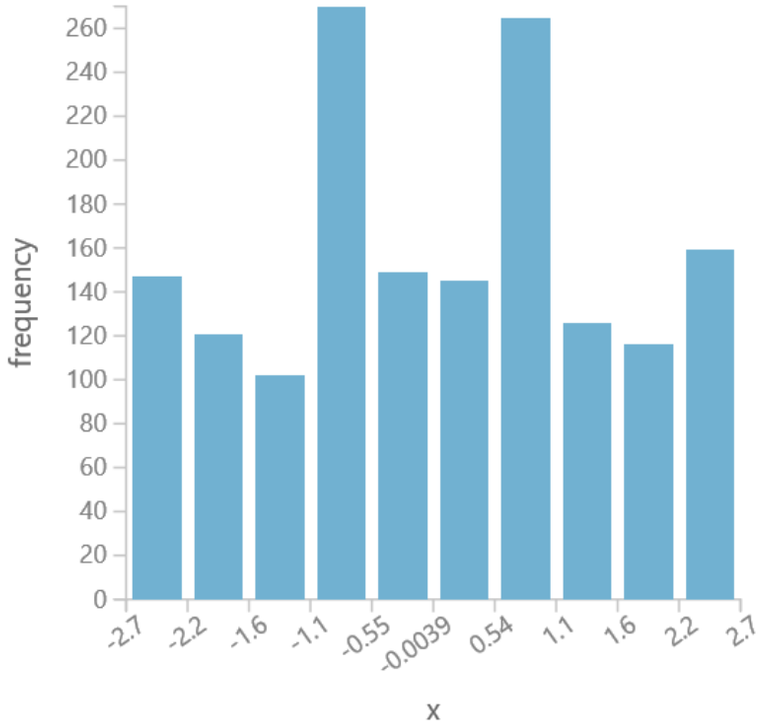
view as

x	y	l
0.756759	-1.373646	S
-1.071171	0.934892	L
0.301784	-1.447297	S
-0.071593	2.472344	C
0.470462	-1.388383	S
0.544586	-1.359058	S
0.689908	-1.280118	S
-0.317403	-1.500969	S
2.268651	0.537407	C
1.111744	-1.283004	S
-0.032994	-1.350037	S
-0.597814	-1.441323	S
1.141835	-2.325779	C
0.114728	-1.605493	S

x

Histogram

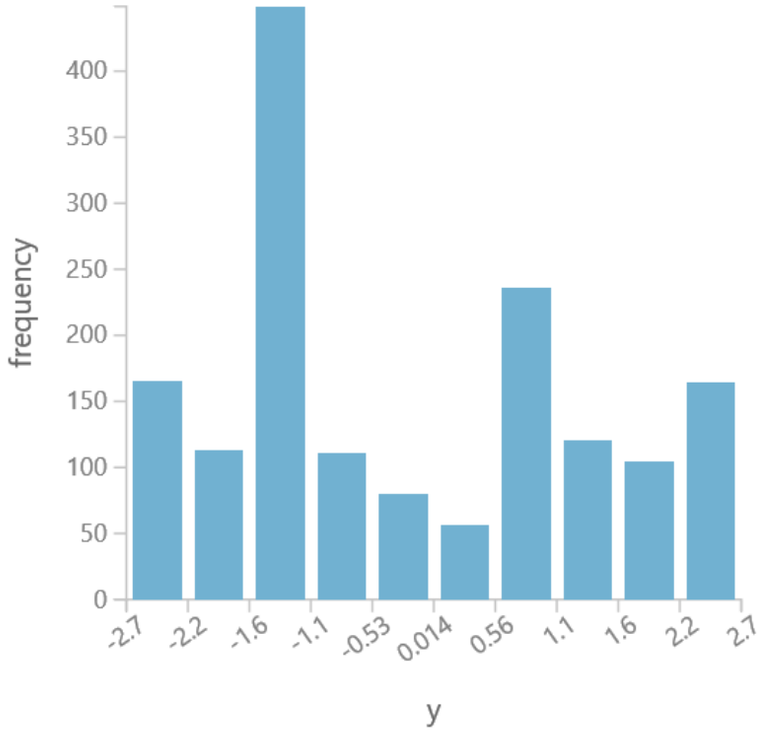
compare to None



y

Histogram

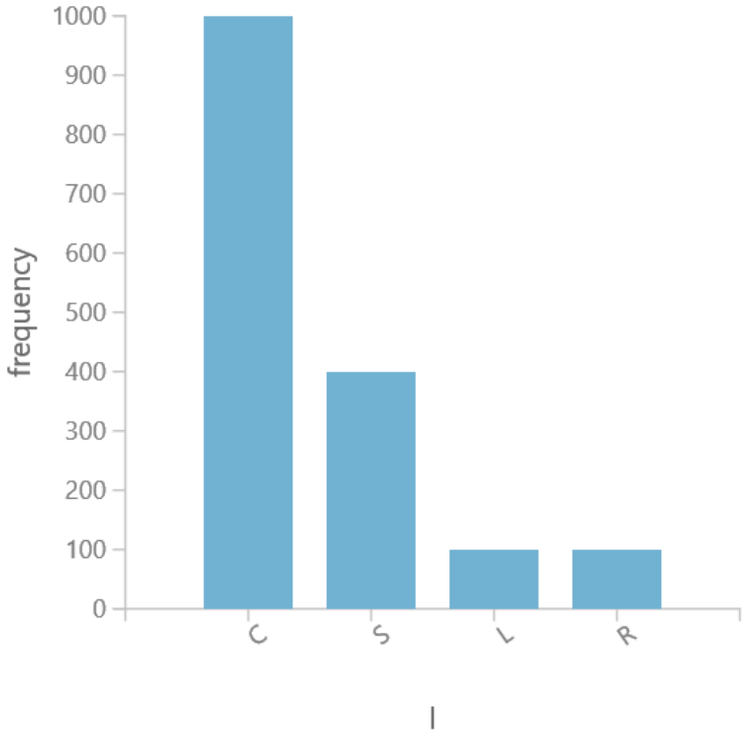
compare to None



l

Histogram

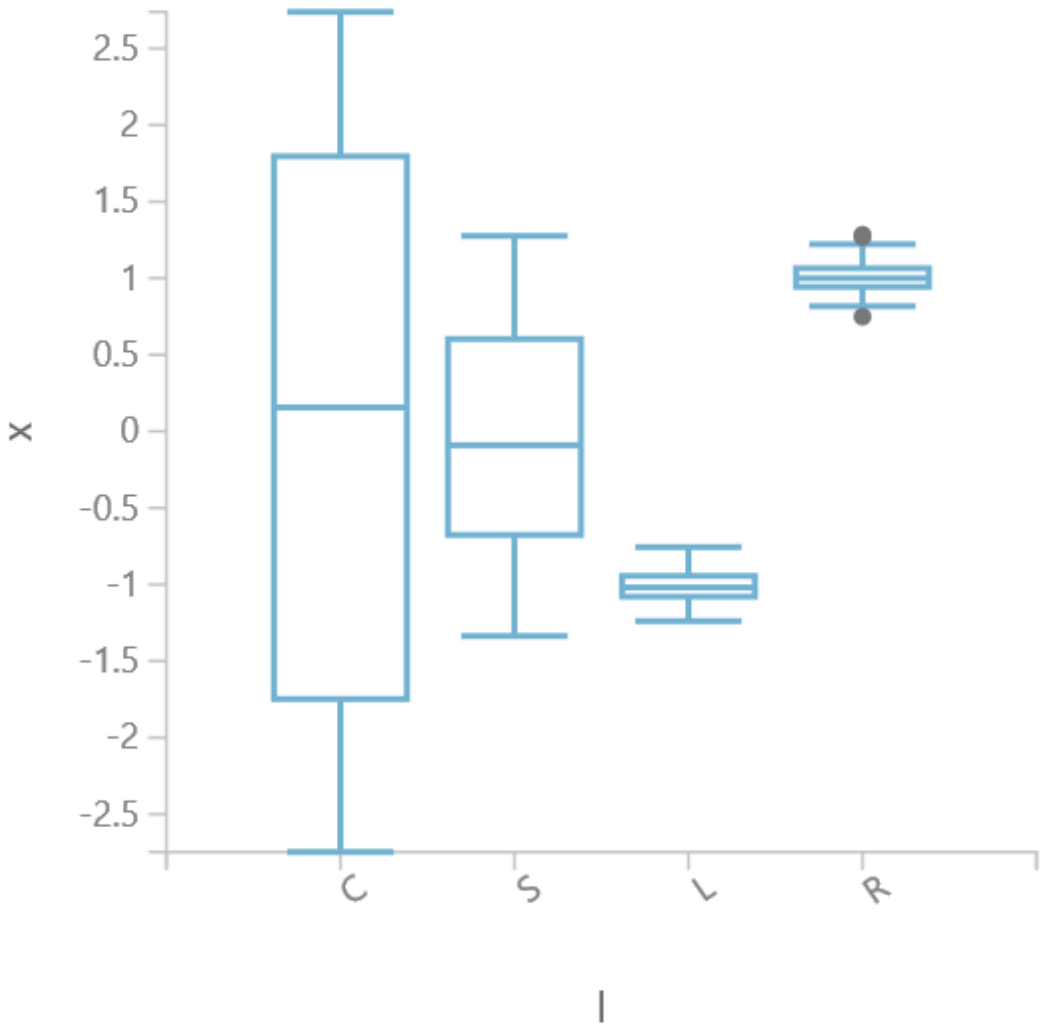
compare to None



x

MultiboxPlot

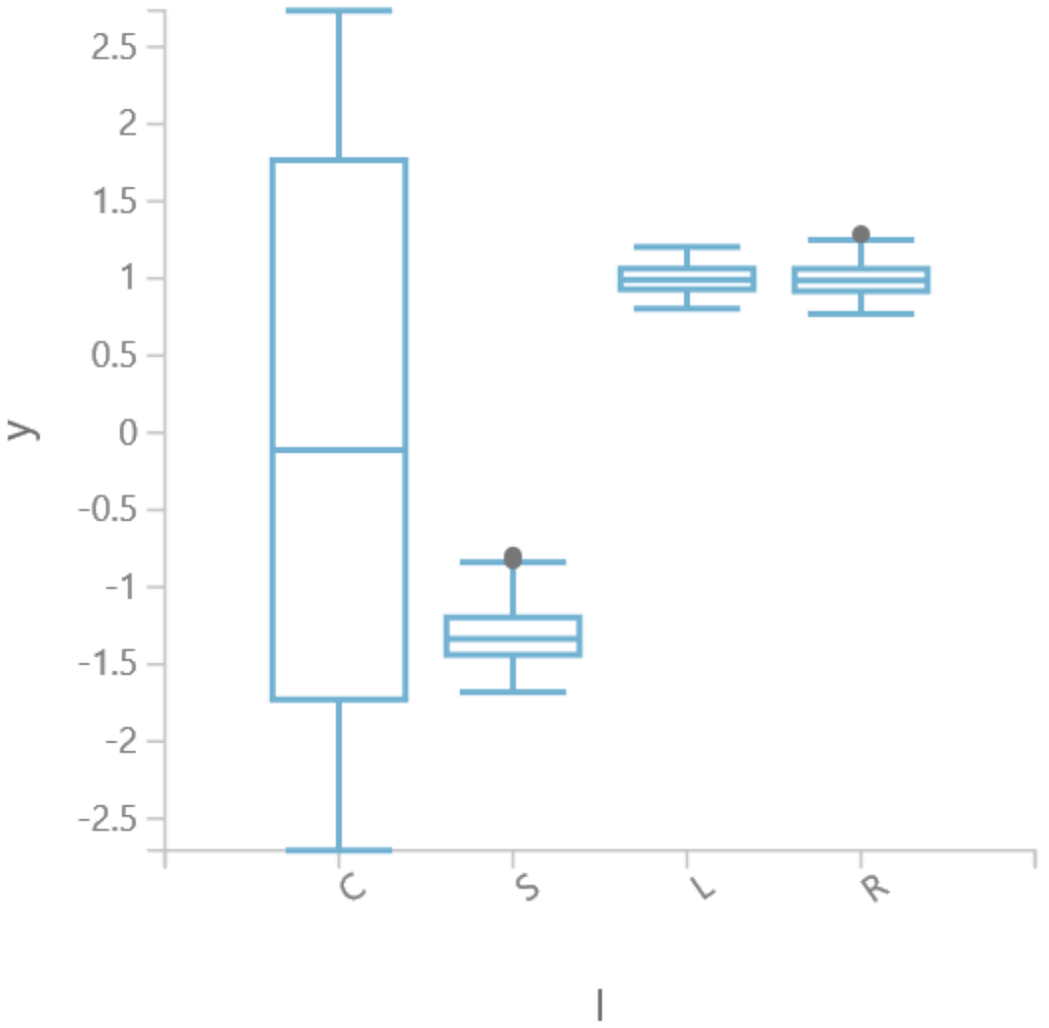
compare to l

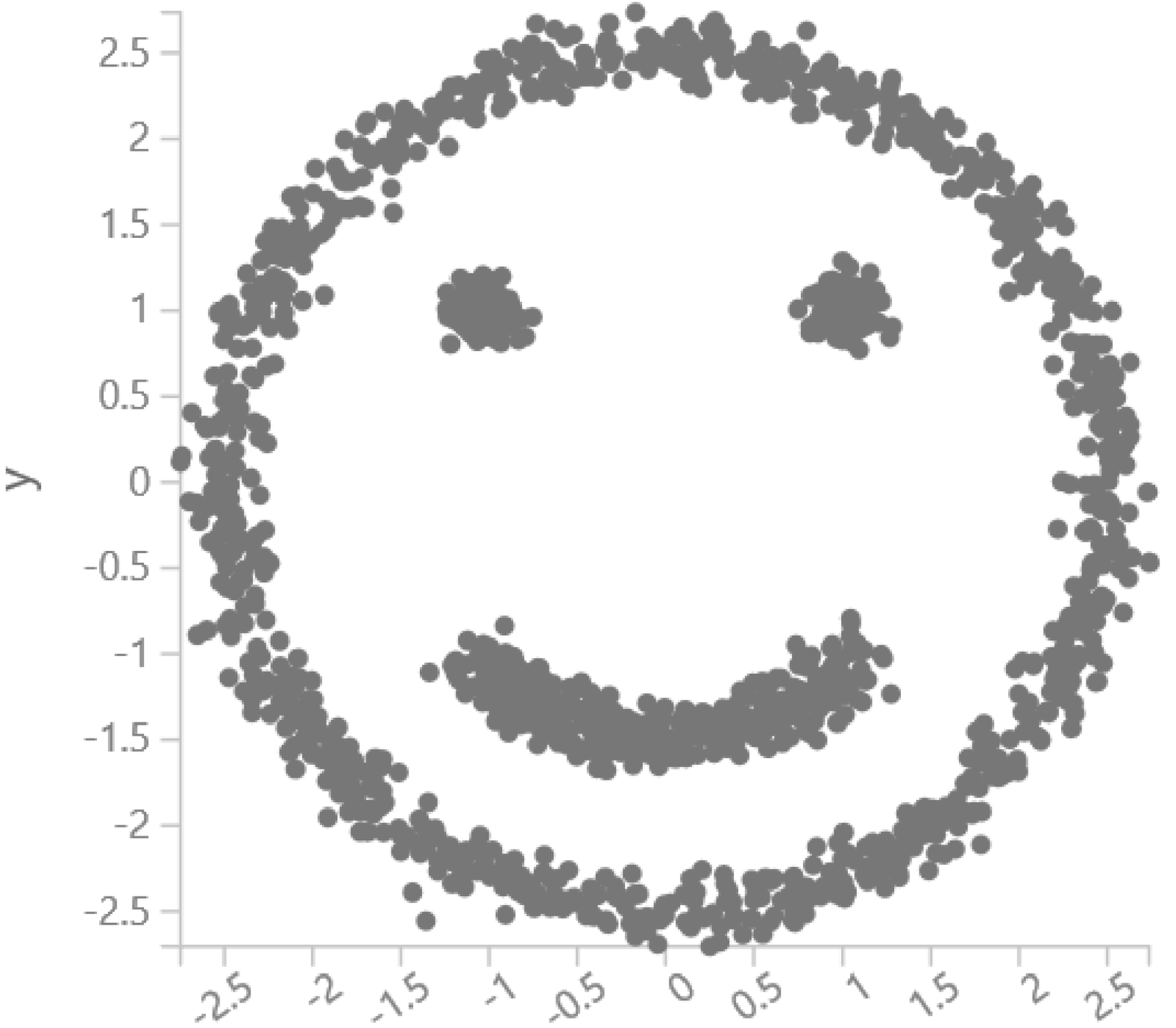


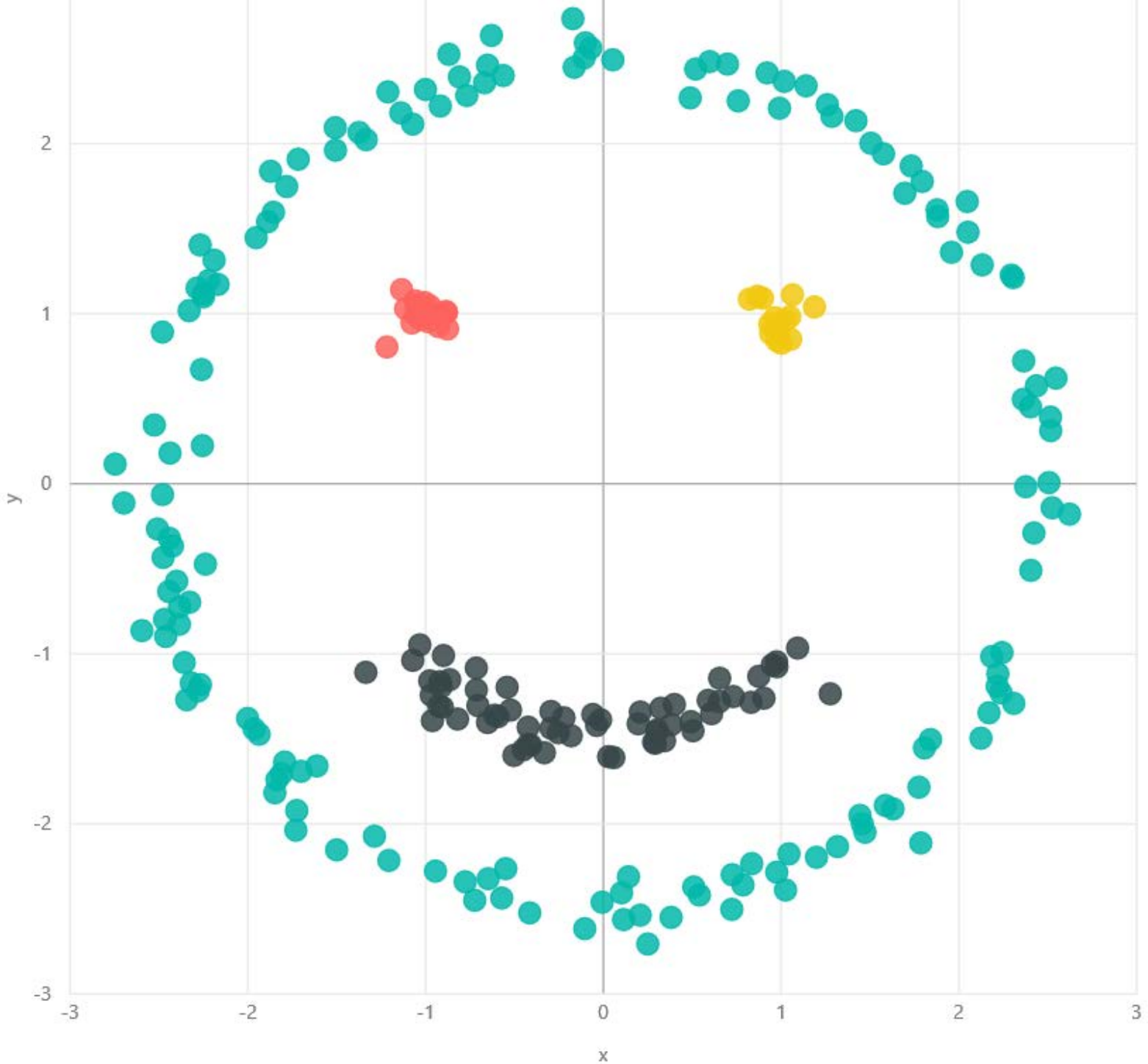
y

MultiboxPlot

compare to l




y
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

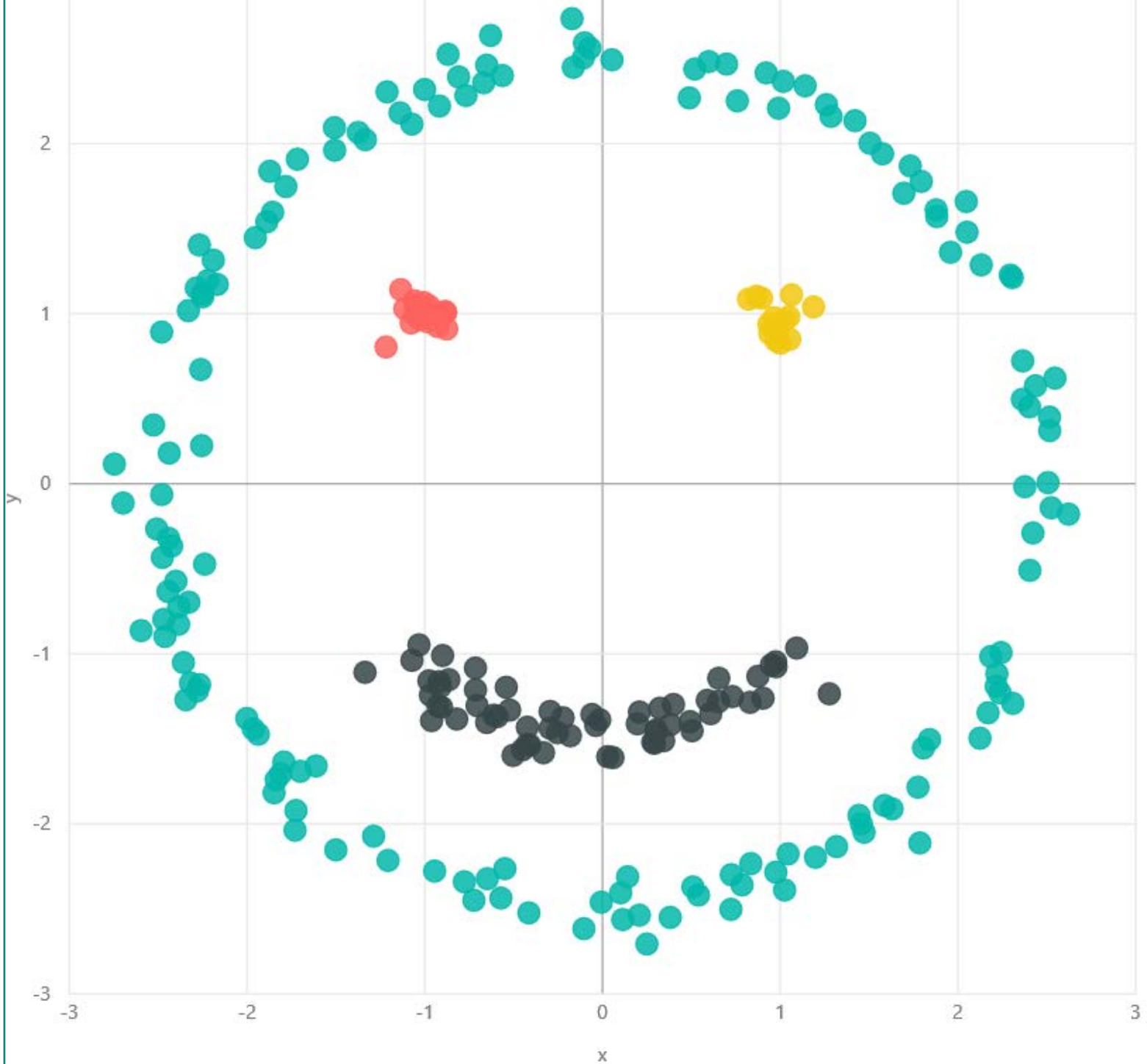

2
1
0
-1
-2
-3
y
-3
-2
-1
0
1
2
3
x

rows 1600

columns 3

view as

x	y	l
0.756759	-1.373646	S
-1.071171	0.934892	L
0.301784	-1.447297	S
-0.071593	2.472344	C
0.470462	-1.388383	S
0.544586	-1.359058	S
0.689908	-1.280118	S
-0.317403	-1.500969	S
2.268651	0.537407	C
1.111744	-1.283004	S
-0.032994	-1.350037	S
-0.597814	-1.441323	S
1.141835	-2.325779	C
0.114728	-1.605493	S
-0.961781	-1.3926	S



Data quality

Missing values

Wrong values

Collection flaws

Feature dependence/leakage

Misinterpretation

Know your domain.

Missing values

There are lots of options

Column mean
Column median
Interpolation
Imputation
Not missing at random
Rank order
Missing means zero
Missing means not applicable
Missing means applicable, but lacking information
Mostly empty columns and rows
Multiple representations for “missing”

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
								
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	48	17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27		118	9		shrimp	3	None
Wanda	19		52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	48	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

median fill

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	46	17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27		118	9		shrimp	3	None
Wanda	19		52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	46	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

mean fill

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	46	17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83	12	77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7			2	shrimp		empty
Natasha	26	4			3			_
Carol	46	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

interpolation

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27	250	1	5	shrimp		Pepper
Donald	67	25	86			beef		Jane
Henry	69	21	95			chicken	62	Janet
Janet	62	21	110			beef		Henry
Nick	46	17	250	4				
Bruce	37	14	63		1	veggie		NA
Steve	83	12	77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	46	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

not missing at random

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27	78	1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	34	17	60					
Bruce	37	14	63			veggie		NA
Steve	83	16	77			chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	27	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

imputation (MICE)

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27	250	1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	46	17	250	4				
Bruce				12	1	veggie		NA
Steve				7	1	chicken		n/a
Clint				9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	46	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

missing rank order

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact	parking_space_IsMissing
Tony	48	27	250	1	5	shrimp		Pepper	false
Donald	67	25	86	10	2	beef		Jane	false
Henry	69	21	95	6	1	chicken	62	Janet	false
Janet	62	21	110	3	1	beef		Henry	false
Nick	46	17	250	4					false
Bruce	37	14	63	-99	1				true
Steve	83	12	77	7	1	chicken		n/a	false
Clint	27	9	118	9		shrimp	3	None	false
Wanda	19	7	52	2	2	shrimp		empty	false
Natasha	26	4	162	5	3			_	false
Carol	46	3	127	11	1	veggie	1	""""	false
Mandy	44	2	68	8	1	chicken		null	false

no information

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact	parking_space_IsMissing
Tony	48	27	250	1	5	shrimp		Pepper	false
Donald	67	25	86	10	2	beef		Jane	false
Henry	69	21	95	6	1	chicken	62	Janet	false
Janet	62	21	110	3	1	beef		Henry	false
Nick	46	17	250	4	0				false
Bruce	37	14	63	-99	1				true
Steve	83	12	77	7	1				false
Clint	27	9	118	9	0				false
Wanda	19	7	52	2	2	shrimp		empty	false
Natasha	26	4	162	5	3			_	false
Carol	46	3	127	11	1	veggie	1	""""	false
Mandy	44	2	68	8	1	chicken		null	false

zero replacement

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact	parking_space_IsMissing	attending_party_IsMissing
Tony	48	27	250	1	5	shrimp		Pepper	false	false
Donald	67	25	86	10	2	beef		Jane	false	false
Henry	69	21	95	6	1	chicken	62	Janet	false	false
Janet	62	21	110	3	1	beef		Henry	false	false
Nick	46	17	250	4	0					true
Bruce	37	14	63	-99	1					false
Steve	83	12	77	7	1					false
Clint	27	9	118	9	0					true
Wanda	19	7	52	2	2	shrimp		empty	false	false
Natasha	26	4	162	5	3			_	false	false
Carol	46	3	127	11	1	veggie	1	""""	false	false
Mandy	44	2	68	8	1	chicken		null	false	false

zero replacement with missing flag

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact	parking_space_IsMissing	attending_party_IsMissing	emergency_contact (2)
Tony	48	27	250	1	5	shrimp		Pepper	false	false	present
Donald	67	25	86	10	2	beef		Jane	false	false	present
Henry	69	21	95	6	1	chicken	62	Janet	false	false	present
Janet	62	21						Henry	false	false	present
Nick	46	17						no	false	true	absent
Bruce	37	14						NA	true	false	absent
Steve	83	12						n/a	false	false	absent
Clint	27	9						None	false	true	absent
Wanda	19	7						empty	false	false	absent
Natasha	26	4						_	false	false	absent
Carol	46	3	127	11	1	veggie	1	""""	false	false	absent
Mandy	44	2	68	8	1	chicken		null	false	false	absent

multiple representations for “missing”

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact	parking_space_IsMissing	attending_party_IsMissing
Tony	48	27	250	1	5	shrimp		Pepper	false	false
Donald	67	25	86	10	2	beef		Jane	false	false
Henry	69	21	95	6	1	chicken	62	Janet	false	false
Janet	62	21				beef		Henry	false	false
Nick	46	17						no	false	true
Bruce	37	14				veggie		NA	true	false
Steve	83	12				chicken		n/a	false	false
Clint	27	9	118	9	0	shrimp	3	None	false	true
Wanda	19	7	52	2	2	shrimp		empty	false	false
Natasha	26	4	162	5	3			_	false	false
Carol	46	3	127	11	1	veggie	1	""""	false	false
Mandy	44	2	68	8	1	chicken		null	false	false

drop mostly empty columns

drop rows missing critical values

Column 0	age	years_seniority	income	parking_space	attending_party	entree	emergency_contact	parking_space_IsMissing	attending_party_IsMissing	emergency_con (2)
					5	shrimp	Pepper	false	false	present
					2	beef	Jane	false	false	present
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	chicken	Janet	false	false	present
Janet	62	21	110	3	1	beef	Henry	false	false	present
Nick	46	17	250	4	0		no	false	true	absent
Bruce	37	14	63	-99	1	veggie	NA	true	false	absent
Steve	83	12	77	7	1	chicken	n/a	false	false	absent
Clint	27	9	118	9	0	shrimp	None	false	true	absent
Wanda	19	7	52	2	2	shrimp	empty	false	false	absent
Natasha	26	4	162	5	3		_	false	false	absent
Carol	46	3	127	11	1	veggie	""""	false	false	absent
Mandy	44	2	68	8	1	chicken	null	false	false	absent

Column 0	age	years_seniority	income	parking_space	attending_party	entree	emergency_contact	parking_space_IsMissing	attending_party_IsMissing	emergency_cont (2)
										
Tony	48	27	250	1	5	shrimp	Pepper	false	false	present
Donald	67	25	86	10	2	beef	Jane	false	false	present
Henry	69	21	95	6	1	chicken	Janet	false	false	present
Janet	62	21	110	3	1	beef	Henry	false	false	present
Bruce	37	14	63	-99	1	veggie	NA	true	false	absent
Steve	83	12	77	7	1	chicken	n/a	false	false	absent
Clint	27	9	118	9	0	shrimp	None	false	true	absent
Wanda	19	7	52	2	2	shrimp	empty	false	false	absent
Carol	46	3	127	11	1	veggie	""""	false	false	absent
Mandy	44	2	68	8	1	chicken	null	false	false	absent

Feature Engineering

Domain knowledge
Combining features
Deep learning
Ignoring irrelevant features

Know your question.

rows	columns
65670	3

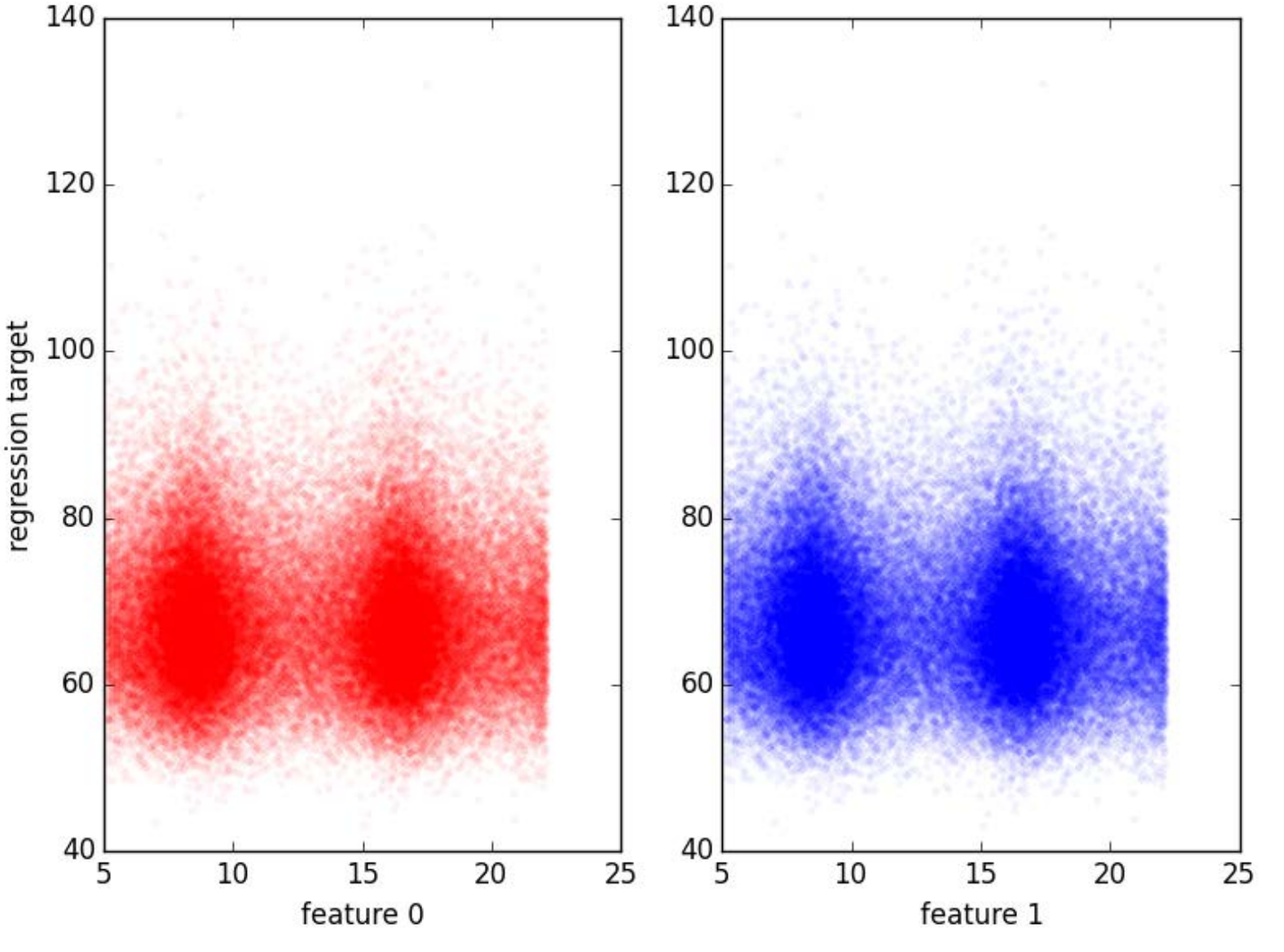
view as

0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809

rows	columns
65670	3

view as

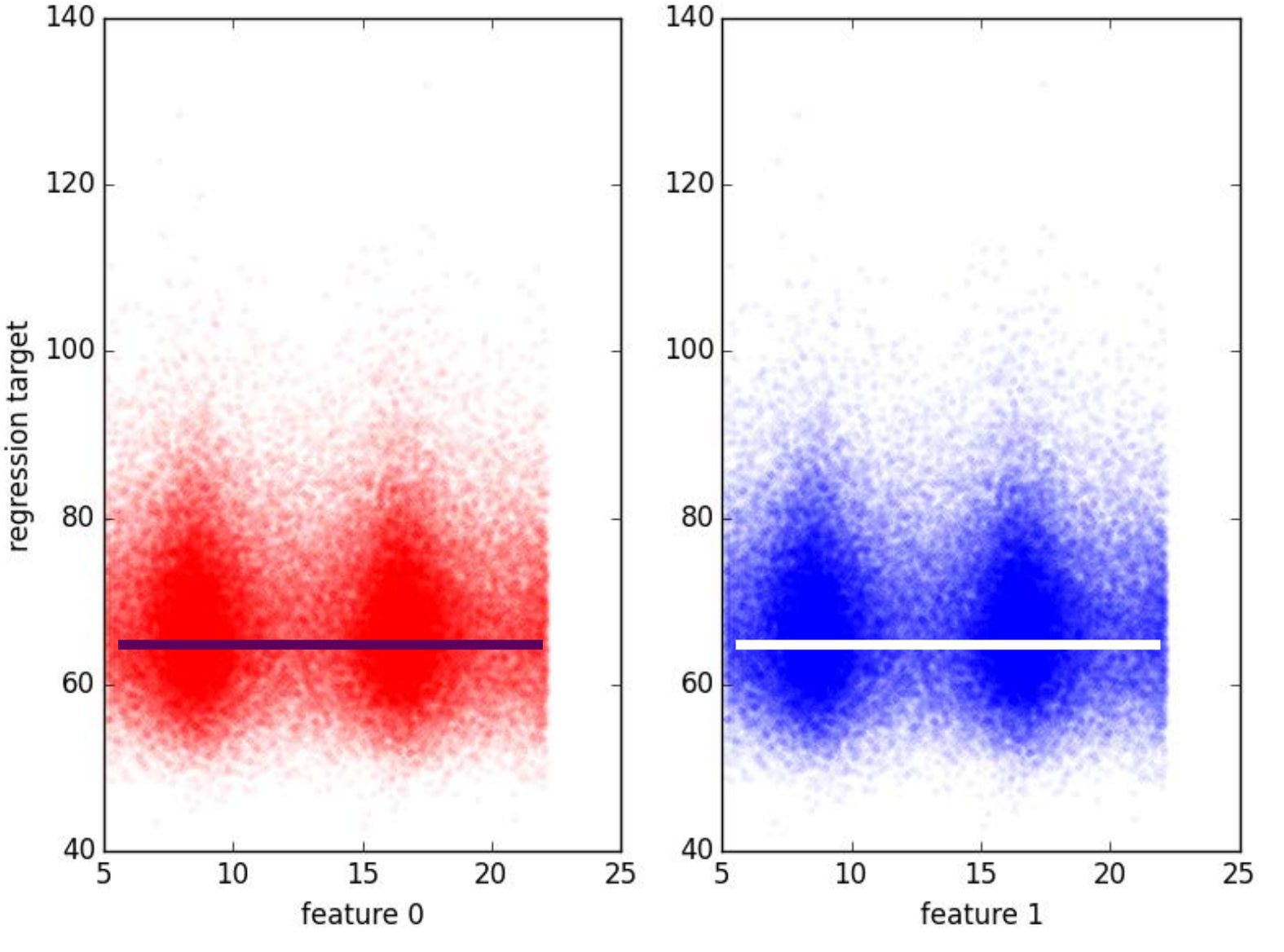
0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809



rows	columns
65670	3

view as

0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809



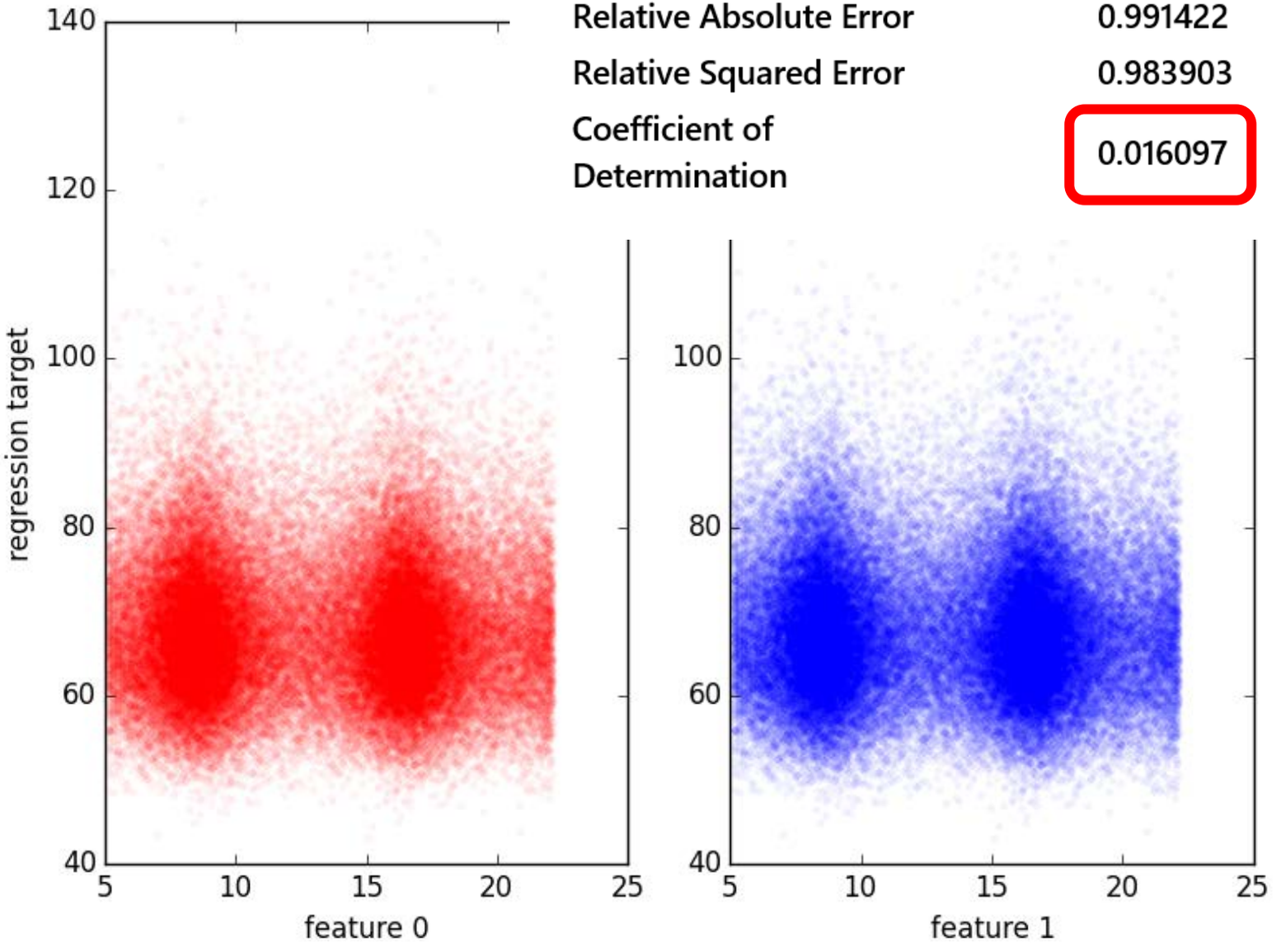
rows	columns
65670	3

view as

0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809

Metrics

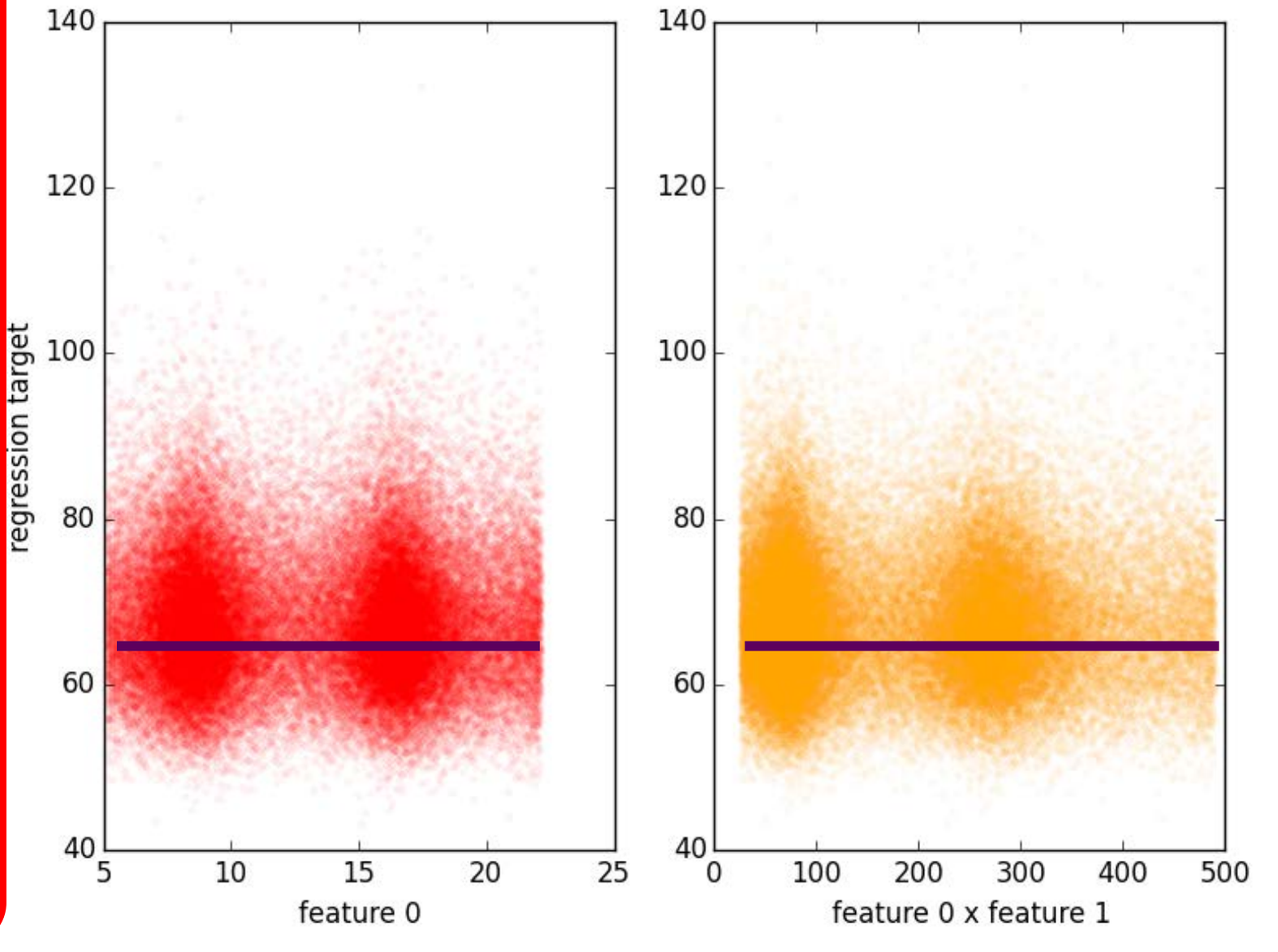
Mean Absolute Error	6.485434
Root Mean Squared Error	8.280206
Relative Absolute Error	0.991422
Relative Squared Error	0.983903
Coefficient of Determination	0.016097



columns

4

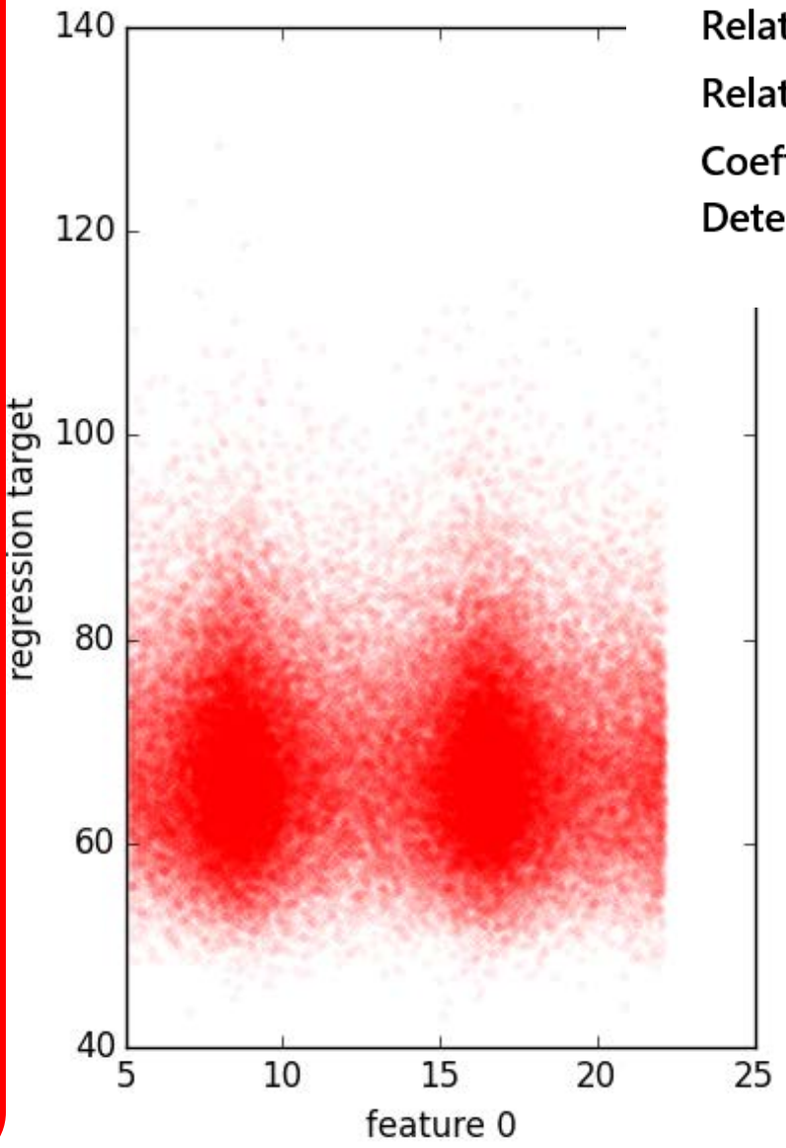
0	1	2	Multiply(1_0)
5.107477	5.135881	60.479023	26.231395
5.113939	5.141432	61.419001	26.292971
5.117143	5.13772	82.774271	26.290449
5.118805	5.145063	62.552338	26.336574
5.119299	5.144294	66.799533	26.335178
5.11949	5.140815	77.870507	26.318351
5.120502	5.147892	64.326006	26.359789
5.121868	5.14889	61.743756	26.371937
5.121949	5.149292	64.493967	26.374413
5.123392	5.148504	69.140338	26.3778
5.124216	5.148921	69.449809	26.384186
5.126409	5.154655	62.028089	26.42487



columns

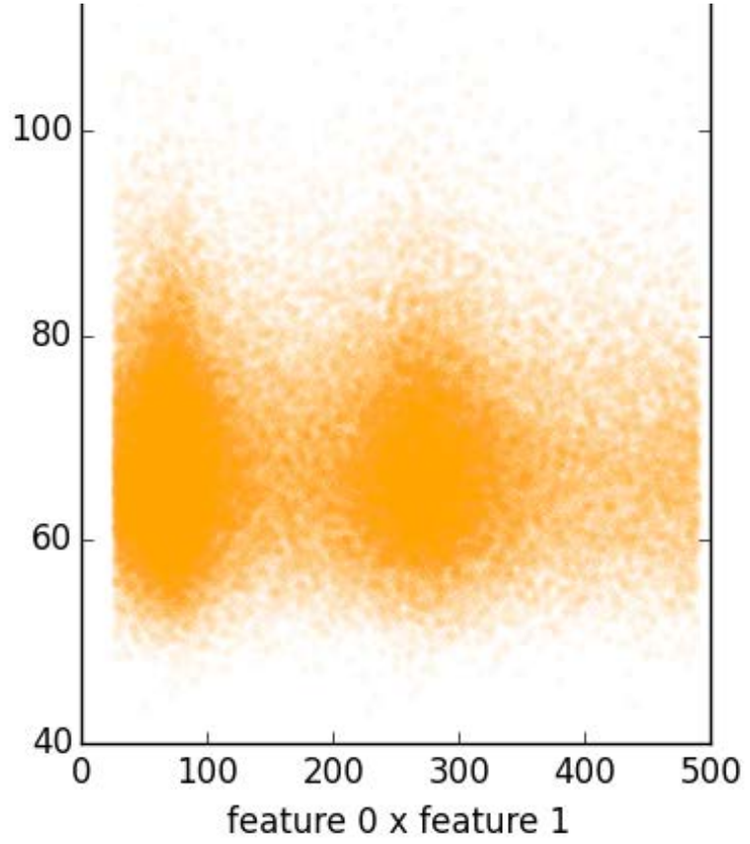
4

0	1	2	Multiply(1_0)
5.107477	5.135881	60.479023	26.231395
5.113939	5.141432	61.419001	26.292971
5.117143	5.13772	82.774271	26.290449
5.118805	5.145063	62.552338	26.336574
5.119299	5.144294	66.799533	26.335178
5.11949	5.140815	77.870507	26.318351
5.120502	5.147892	64.326006	26.359789
5.121868	5.14889	61.743756	26.371937
5.121949	5.149292	64.493967	26.374413
5.123392	5.148504	69.140338	26.3778
5.124216	5.148921	69.449809	26.384186
5.126409	5.154655	62.028089	26.42487



Metrics

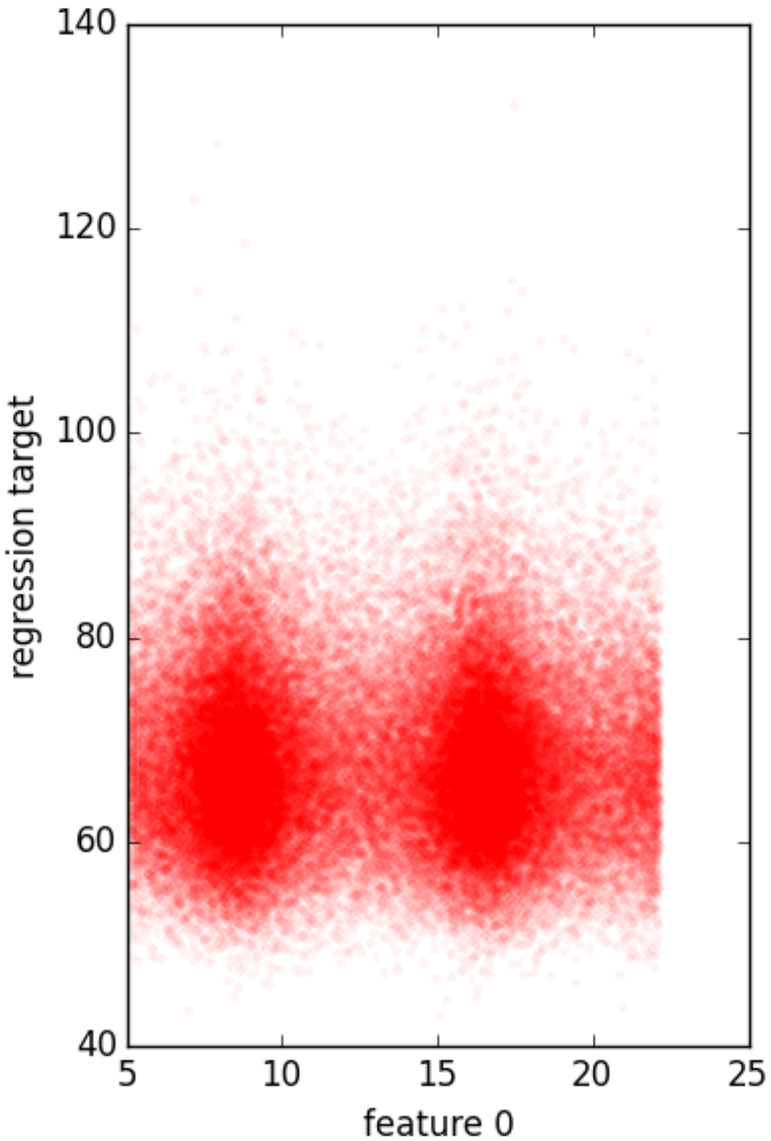
Mean Absolute Error	6.491614
Root Mean Squared Error	8.285875
Relative Absolute Error	0.992366
Relative Squared Error	0.98525
Coefficient of Determination	0.01475

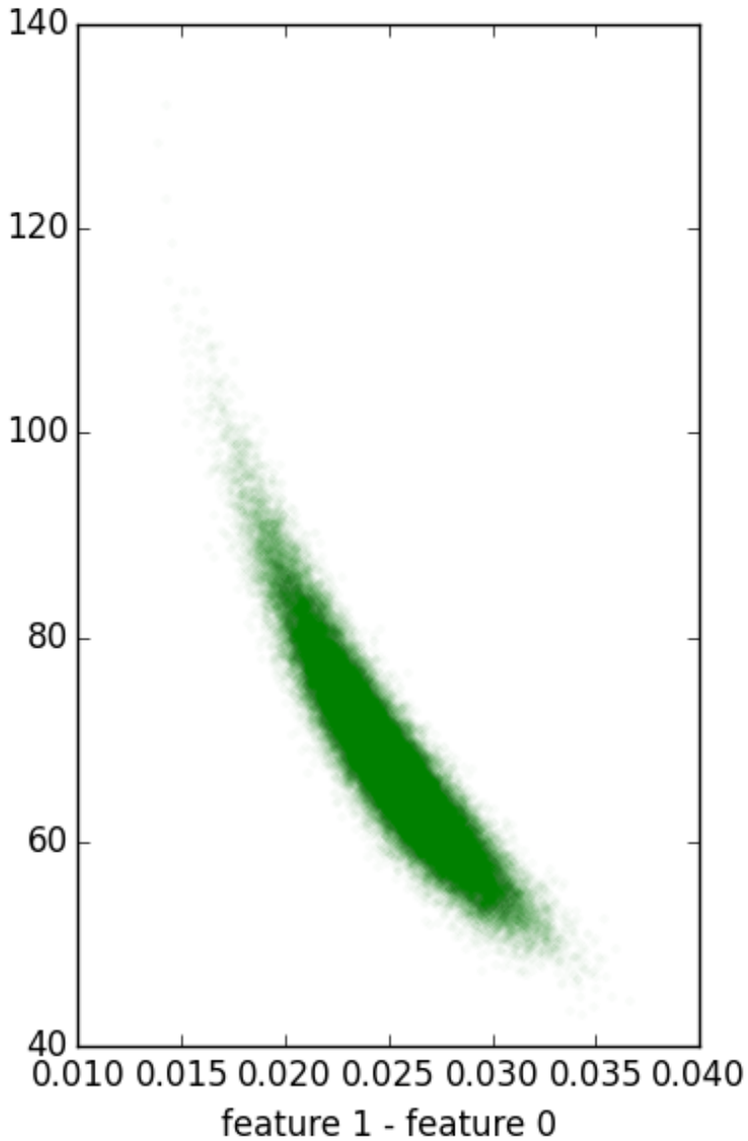


columns

4

0	1	2	Subtract(1_0)
5.107477	5.135881	60.479023	0.028404
5.113939	5.141432	61.419001	0.027493
5.117143	5.13772	82.774271	0.020578
5.118805	5.145063	62.552338	0.026258
5.119299	5.144294	66.799533	0.024995
5.11949	5.140815	77.870507	0.021325
5.120502	5.147892	64.326006	0.02739
5.121868	5.14889	61.743756	0.027022
5.121949	5.149292	64.493967	0.027343
5.123392	5.148504	69.140338	0.025112
5.124216	5.148921	69.449809	0.024705
5.126409	5.154655	62.028089	0.028246

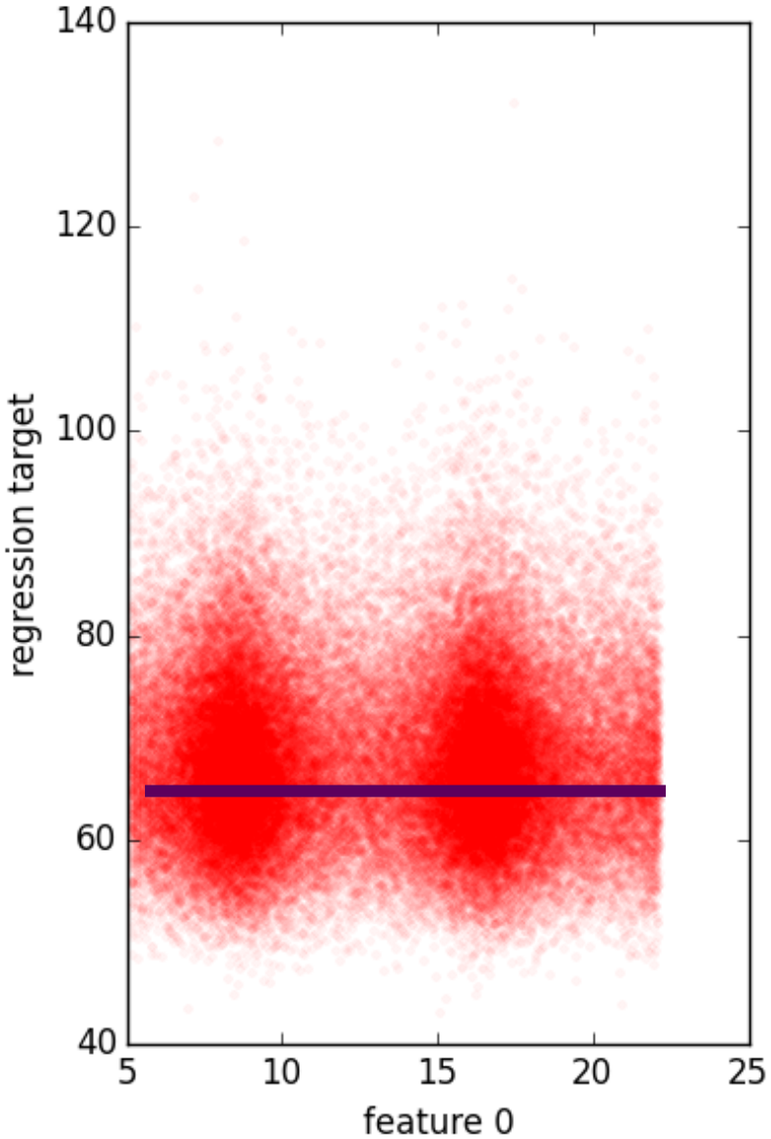


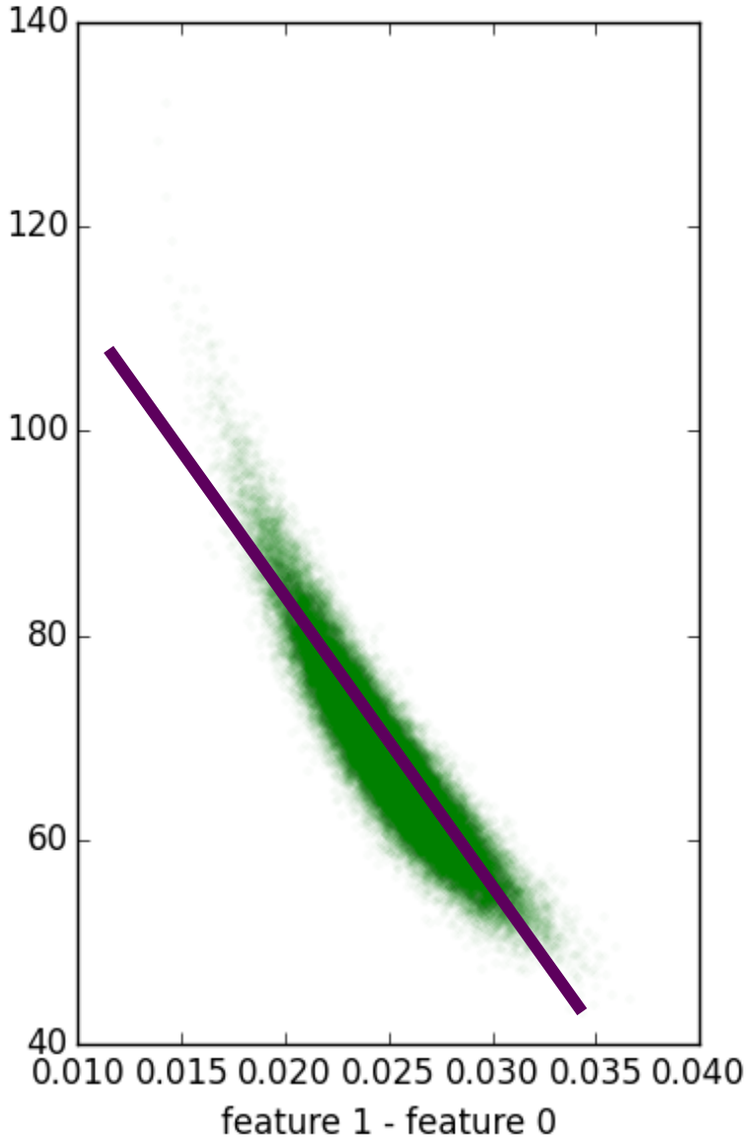


columns

4

0	1	2	Subtract(1_0)
5.107477	5.135881	60.479023	0.028404
5.113939	5.141432	61.419001	0.027493
5.117143	5.13772	82.774271	0.020578
5.118805	5.145063	62.552338	0.026258
5.119299	5.144294	66.799533	0.024995
5.11949	5.140815	77.870507	0.021325
5.120502	5.147892	64.326006	0.02739
5.121868	5.14889	61.743756	0.027022
5.121949	5.149292	64.493967	0.027343
5.123392	5.148504	69.140338	0.025112
5.124216	5.148921	69.449809	0.024705
5.126409	5.154655	62.028089	0.028246

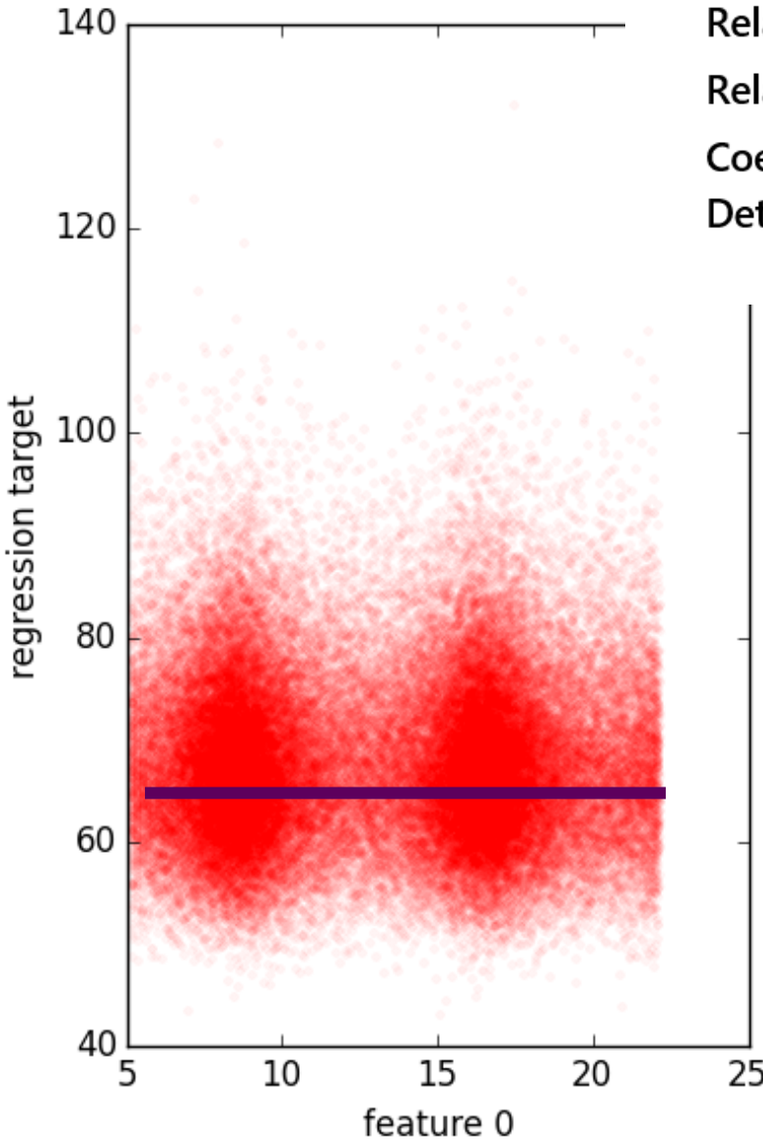




columns

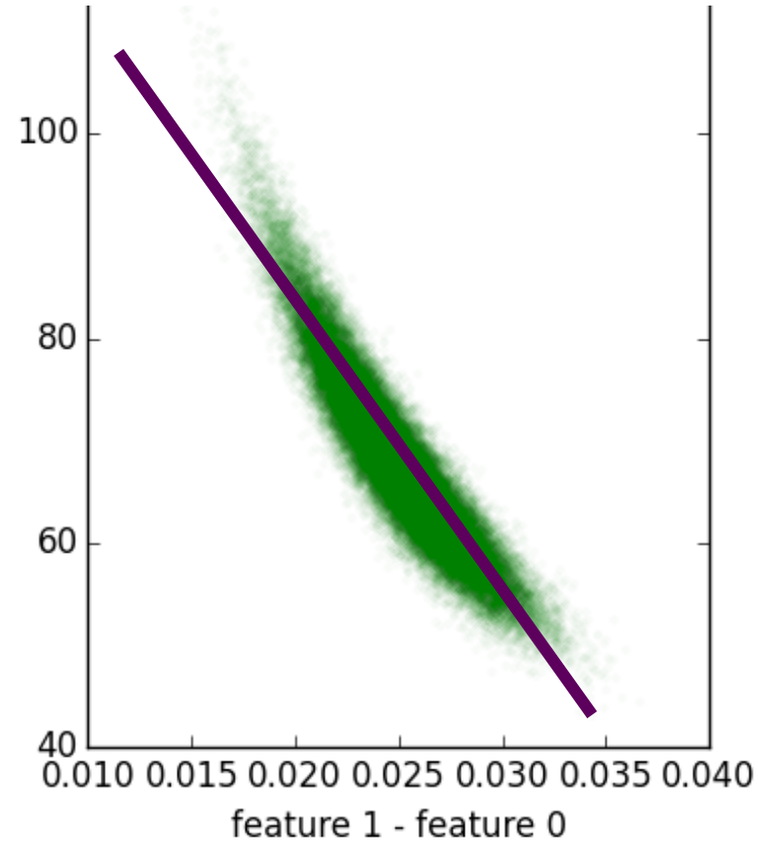
4

0	1	2	Subtract(1_0)
5.107477	5.135881	60.479023	0.028404
5.113939	5.141432	61.419001	0.027493
5.117143	5.13772	82.774271	0.020578
5.118805	5.145063	62.552338	0.026258
5.119299	5.144294	66.799533	0.024995
5.11949	5.140815	77.870507	0.021325
5.120502	5.147892	64.326006	0.02739
5.121868	5.14889	61.743756	0.027022
5.121949	5.149292	64.493967	0.027343
5.123392	5.148504	69.140338	0.025112
5.124216	5.148921	69.449809	0.024705
5.126409	5.154655	62.028089	0.028246



Metrics

Mean Absolute Error	2.243981
Root Mean Squared Error	2.834526
Relative Absolute Error	0.343035
Relative Squared Error	0.1153
Coefficient of Determination	0.8847



Operationalization

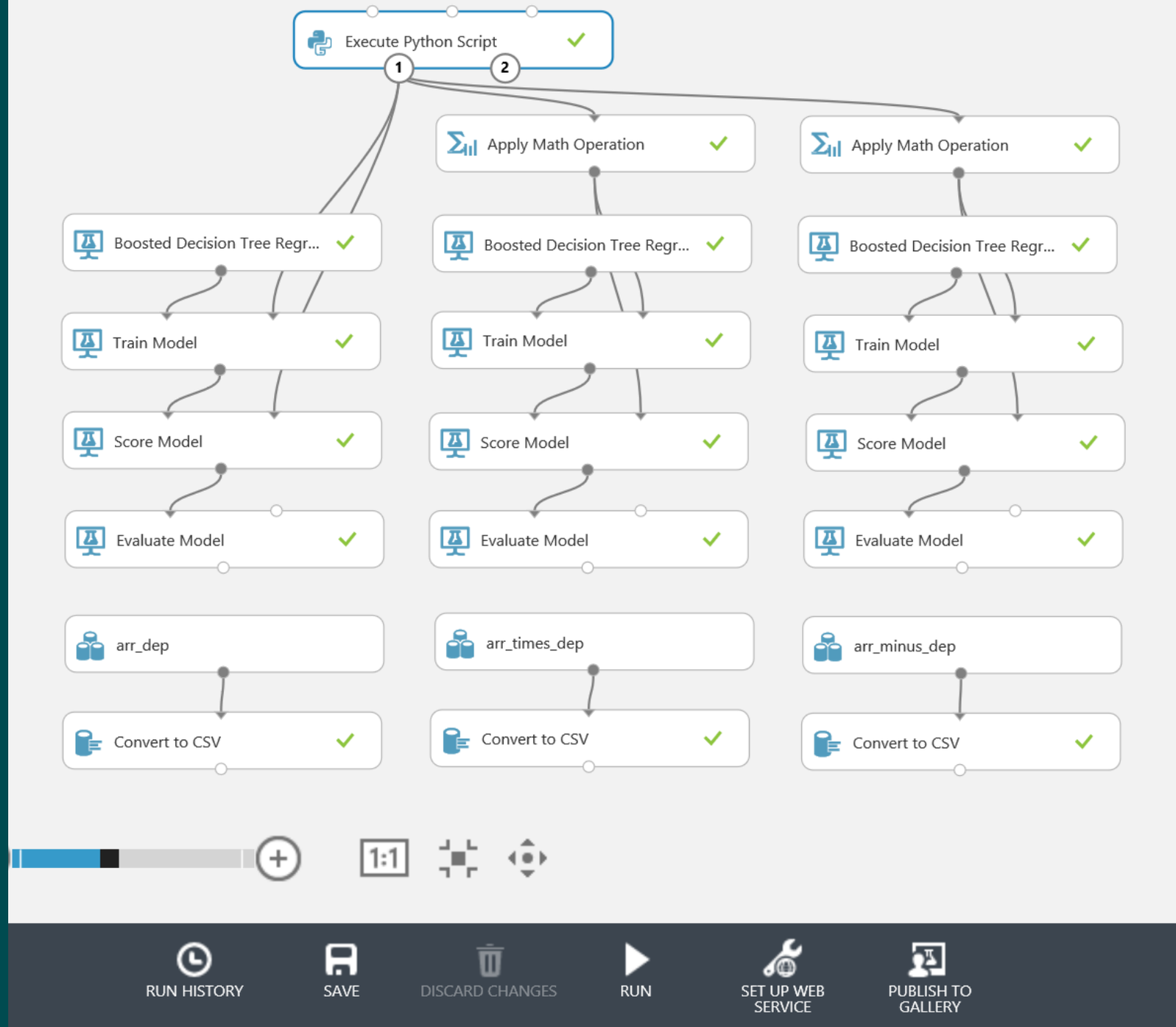
Getting the data to the algorithm

Getting the data cleaned

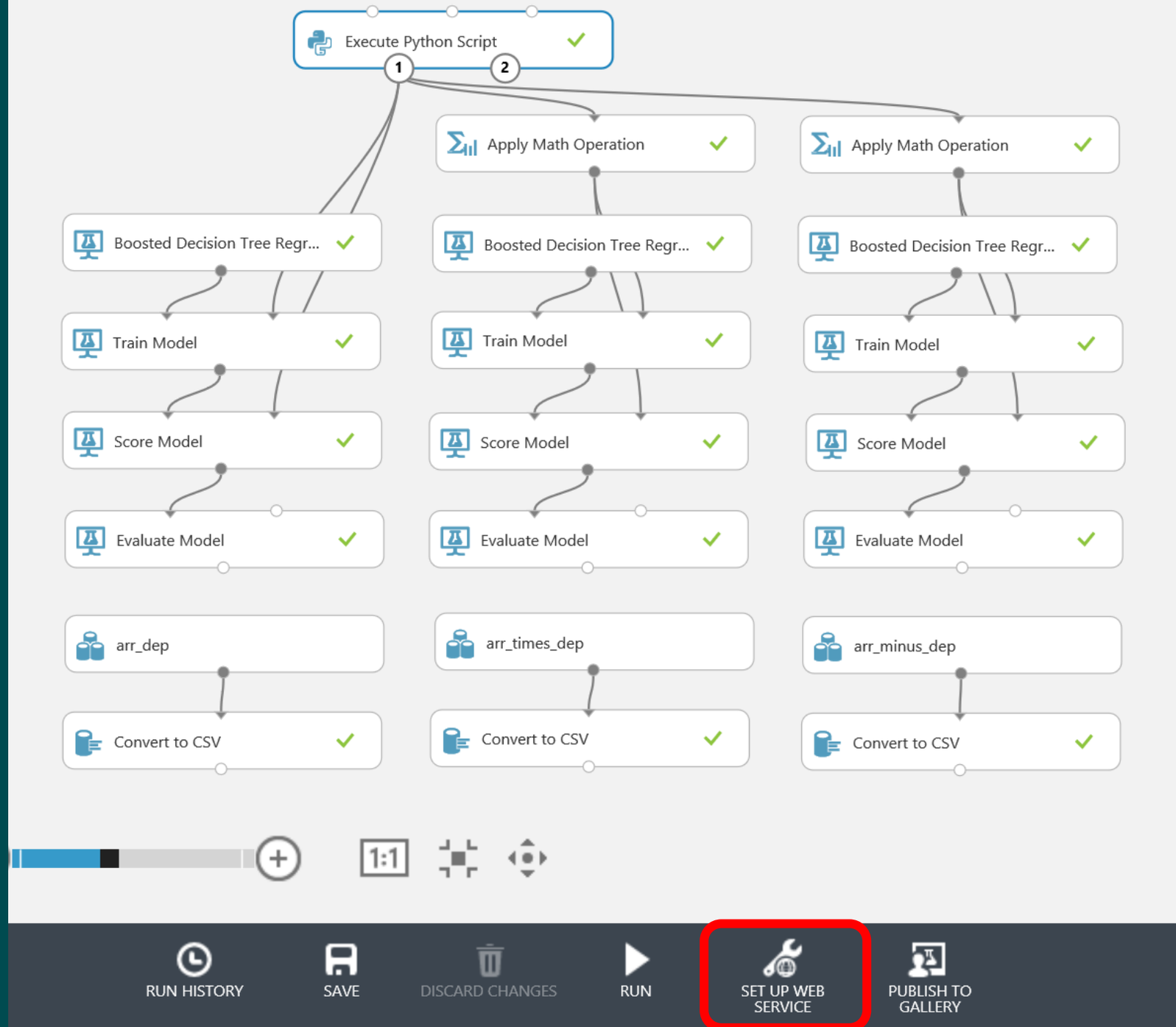
Getting the results to those who need them

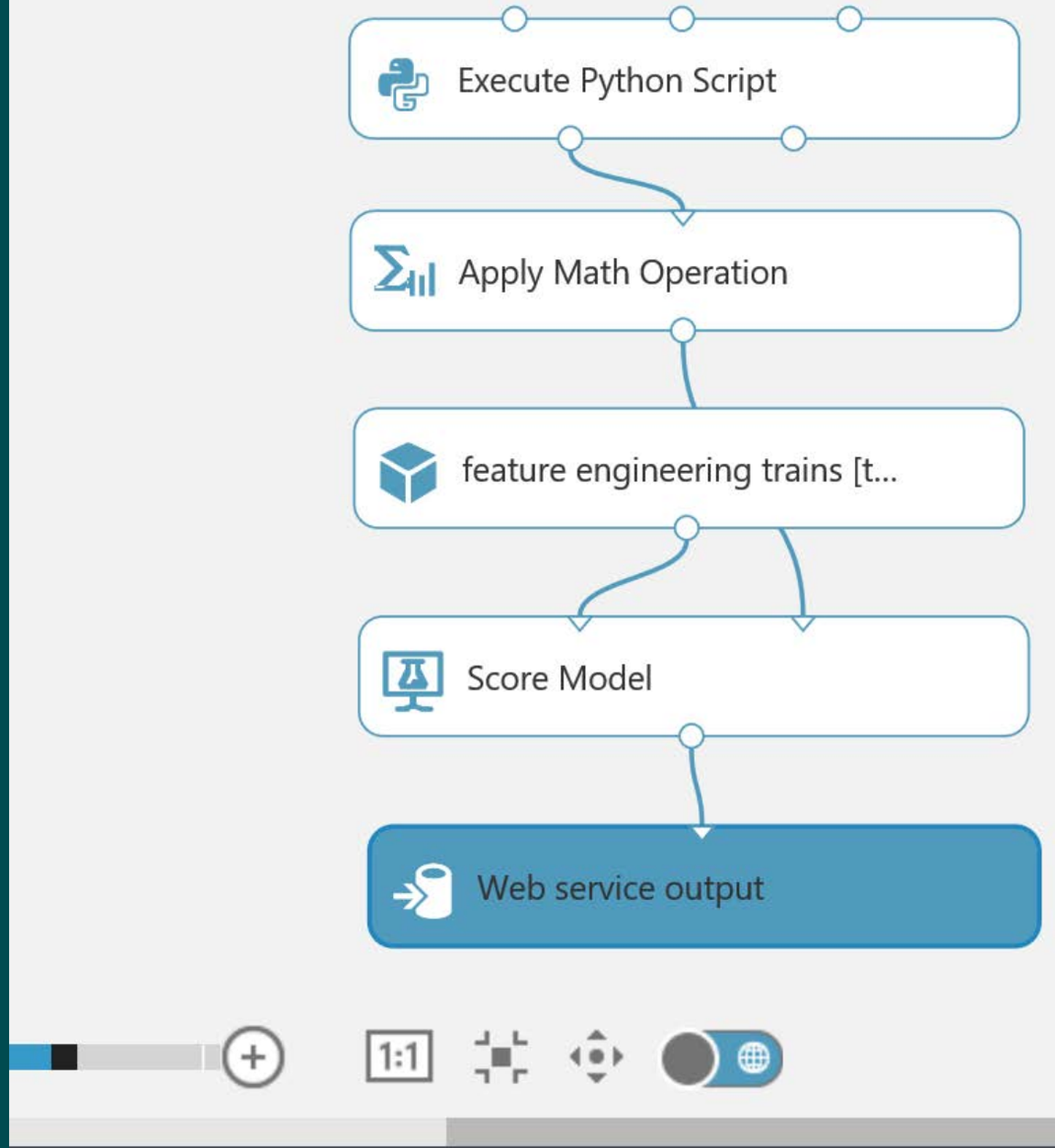
Know how you are going to use the answer.

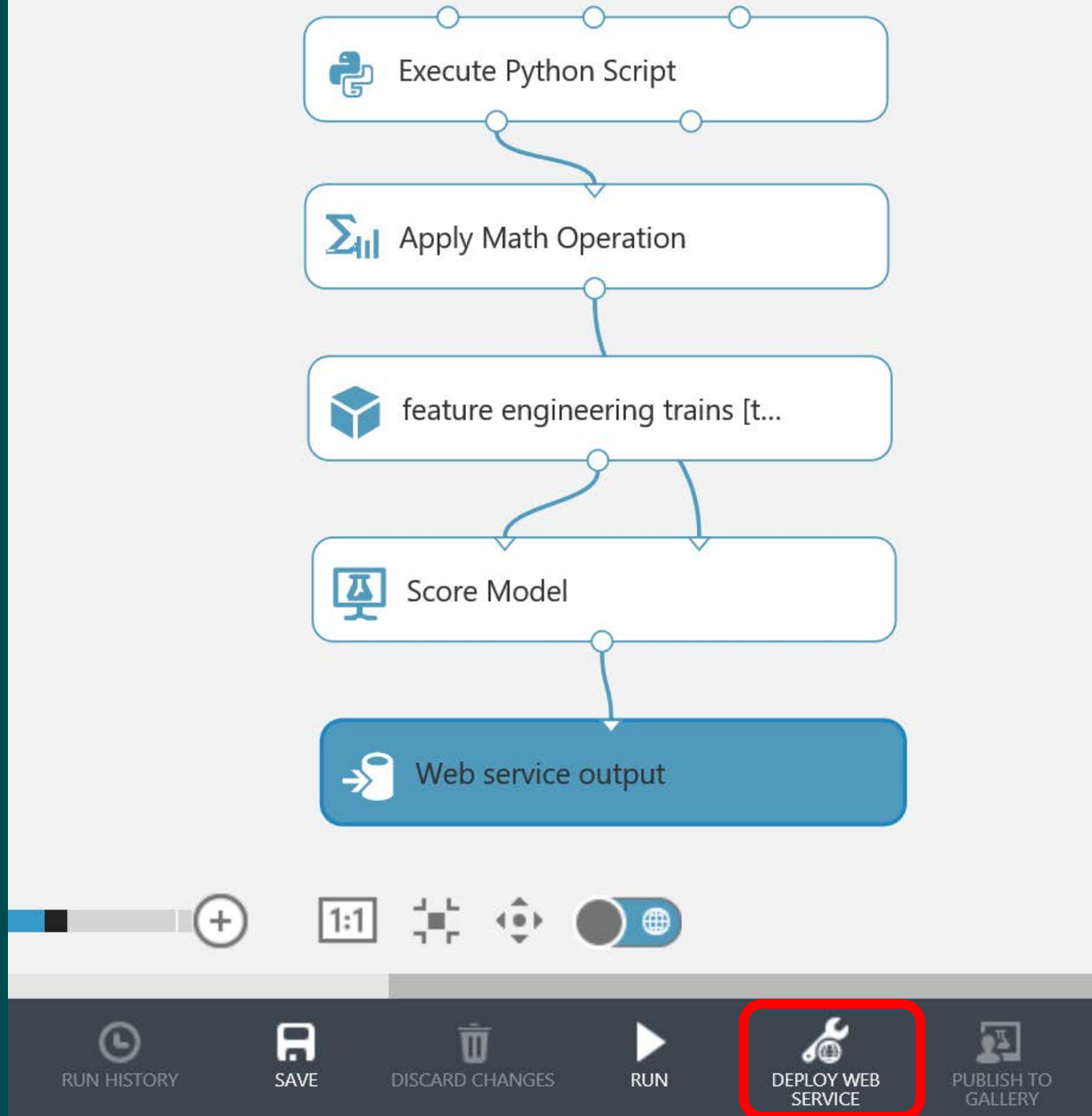
Creating a RESTful API



Creating a RESTful API




Execute Python Script
Apply Math Operation
feature engineering trains [t...
Score Model
Web service output
1:1
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
DEPLOY WEB SERVICE
PUBLISH TO GALLERY


Execute Python Script
Apply Math Operation
feature engineering trains [t...
Score Model
Web service output
1:1
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
DEPLOY WEB SERVICE
PUBLISH TO GALLERY

feature engineering: trains [scoring exp.]

DASHBOARD CONFIGURATION

General

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

oPYrmjj9qpuZlgdezr2PTPifG0D2YJ88v5gUOde2uGCNS9pNl6IoA2QHSEwSdcAUhEfD1QSU6SIzVc1BsqJiQQ==

Default Endpoint

API HELP PAGE	TEST	APPS
REQUEST/RESPONSE	Test	Download Excel Workbook
BATCH EXECUTION		

feature engineering: trains [scoring exp.]

DASHBOARD CONFIGURATION

General

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

oPYrmjj9qpuZlgdezr2PTPifG0D2YJ88v5gUOde2uGCNS9pNl6IoA2QHSEwSdcAUhEfD1QSU6SIzVc1BsqJiQQ==

Default Endpoint

API HELP PAGE	TEST	APPS
REQUEST/RESPONSE	Test	Download Excel Workbook
BATCH EXECUTION		

feature engineering: trains [scoring exp.]

DASHBOARD CONFIGURATION

General

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

oPYrmjj9qpuZlgdezr2PTPifG0D2YJ88v5gUOde2uGCNS9pNl6IoA2QHSEwSdcAUhEfD1QSU6SIzVc1BsqJiQQ==

Default Endpoint

API HELP PAGE	TEST	APPS
REQUEST/RESPONSE	Test	Download Excel Workbook
BATCH EXECUTION		

Request Response API Documentation for feature engineering: tr [Scoring Exp.]

Updated: 10/08/2015 22:03

No description provided for this web service.

- Previous version of this API
- Submit a request
- Output Parameters
- ASP.Net App Template for RRS
- Sample Code

OData Endpoint Address

```
https://ussouthcentral.services.azureml.net/odata/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c89b6a
```

Request

Method	Request URI
POST	`https://ussouthcentral.services.azureml.net/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c89 api-version=2.0&details=true`

Request Response API Documentation for feature engineering: tr [Scoring Exp.]

Updated: 10/08/2015 22:03

No description provided for this web service.

- Previous version of this API
- Submit a request
- Output Parameters
- ASP.Net App Template for RRS
- Sample Code

OData Endpoint Address

```
https://ussouthcentral.services.azureml.net/odata/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c89b6a
```

Request

Method	Request URI
POST	`https://ussouthcentral.services.azureml.net/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c8` `api-version=2.0&details=true`

Know your data.
Know your question.
Know your domain.
Know how you’re going to use the answer.

Smiley data link
Missing values link
Train data link
Collection link