Badge-2 Lab-3 [Linear Regression]

Out date: 20 July 2022
Due date: 24 July 2022 @ 11:59pm

Objective: To apply multi-linear regression to predict a quantitative target variable.

Data:

Horizontal drilling which provides additional exposure of the reservoir to the wellbore (lateral length) is one of the primary drivers that made economic production of oil and gas from the tight shale reservoirs in the US a success, in conjunction with stimulation using hydraulic fracturing.

Objective of the exercise is to understand effect of lateral length of a horizontal well on shale gas production in some of the US shale plays using multiple linear regression analysis. While lateral length would be one of the primary predictors, other variables considered are vertical section length of the horizontal well, proppant volume and fluid volume pumped for stimulating the horizontal section of the well using hydraulic fracturing and type of shale plays.

Data source: National Oil and Gas Gateway (http://www.noggateway.org/explore). From participating states that contribute to this data source, Arkansas and Colorado were identified as states of interest for the project owing to the level of shale gas drilling and production activity in these states over the period 2008-2018 and data availability on Fluids and Proppants.

Variables of interest:

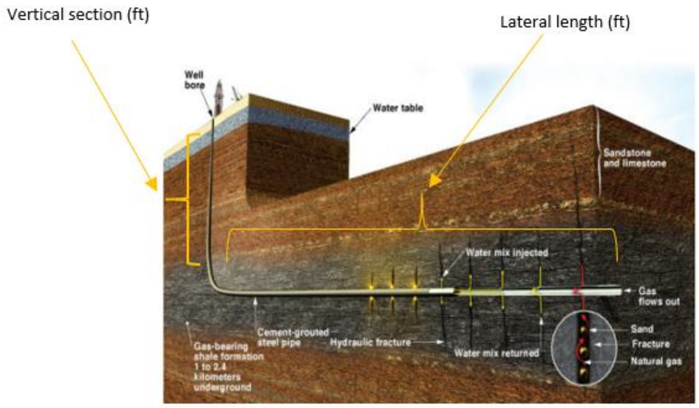


Figure 1: Horizontal well schematic (from The Academy of Medicine, 2017)

Referring to the Figure-1 above, the following were identified as variables of interest after reviewing the dataset:

Response (Target) variable: Max_Gas, Maximum Gas Production (Million Standard Cubic Feet, MSCF)
Since shale gas wells have their peak production in the first 2-3 years after the well is drilled, stimulated and put on production, maximum annual gas production is selected as the best response variable to predict using the predictor variables identified below

Predictor variables:
Lat_Len, Lateral length of the horizontal section (feet)

Measured Depth, Total Depth of the well (**feet**)

Bottom and **Top** Perforation depth (feet)

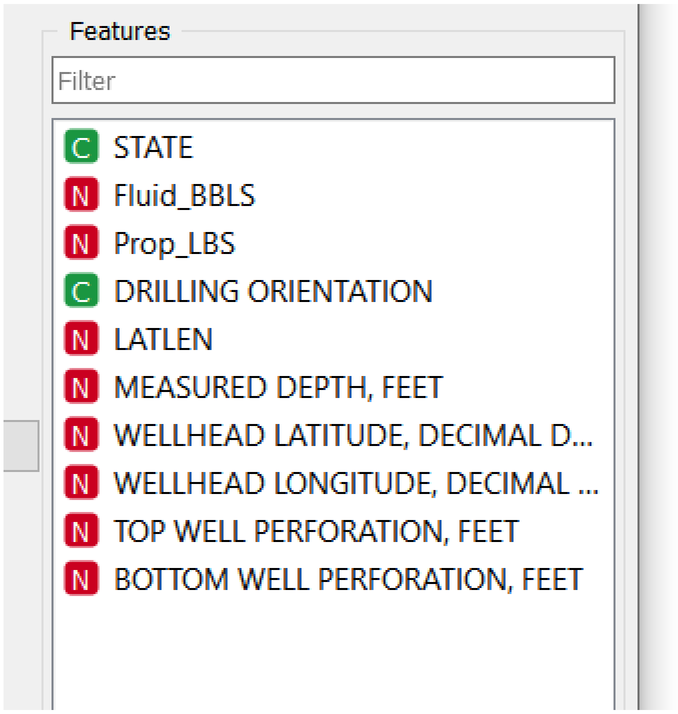
Wellhead **Latidude** and **Longitude** (decimal degrees), surface location of the well

Fluid, Total amount of fracturing fluid (typically water) pumped to create the hydraulic fractures (Barrels)

Proppant, Total amount of proppant (typically sand) pumped along with the fracturing fluid as a slurry to keep the hydraulically created fractures open (Pounds)

ShalePlay (a dummy variable, 0- Fayetteville, Arkansas Shale gas play and 1- Mancos, Colorado Shale gas play)--**STATE**

Drilling Orientation (a dummy variable)



Features

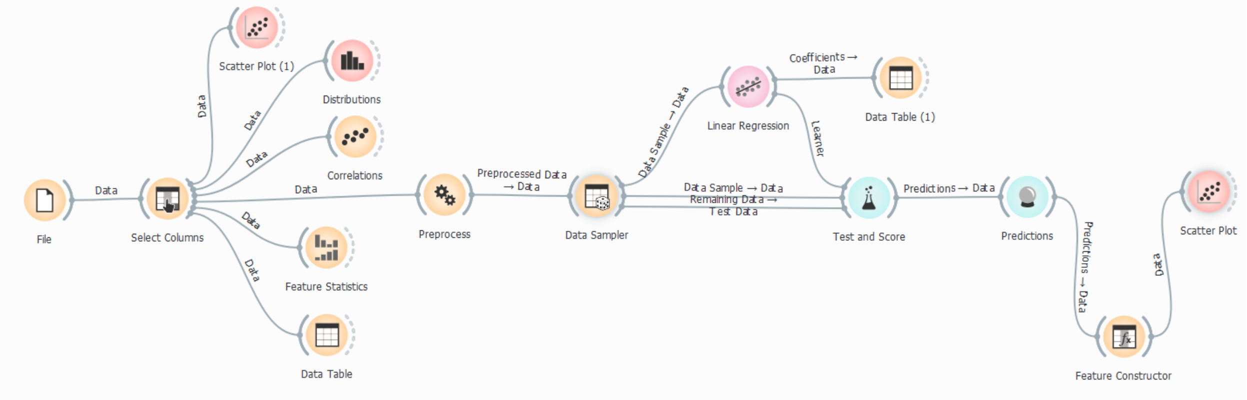
Filter

- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL D...
- N WELLHEAD LONGITUDE, DECIMAL ...
- N TOP WELL PERFORATION, FEET
- N BOTTOM WELL PERFORATION, FEET

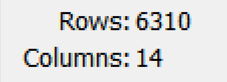
Data: For this lab, please download *production.csv* and *gas_prod_pred_start.ows* to your local folder.

Lab Instructions

1. Open the *gas_prod_pred_start.ows* file in Orange. Your pipeline should look as below. Confirm the *data.xlsx* dataset is loaded by opening the **File** widget.

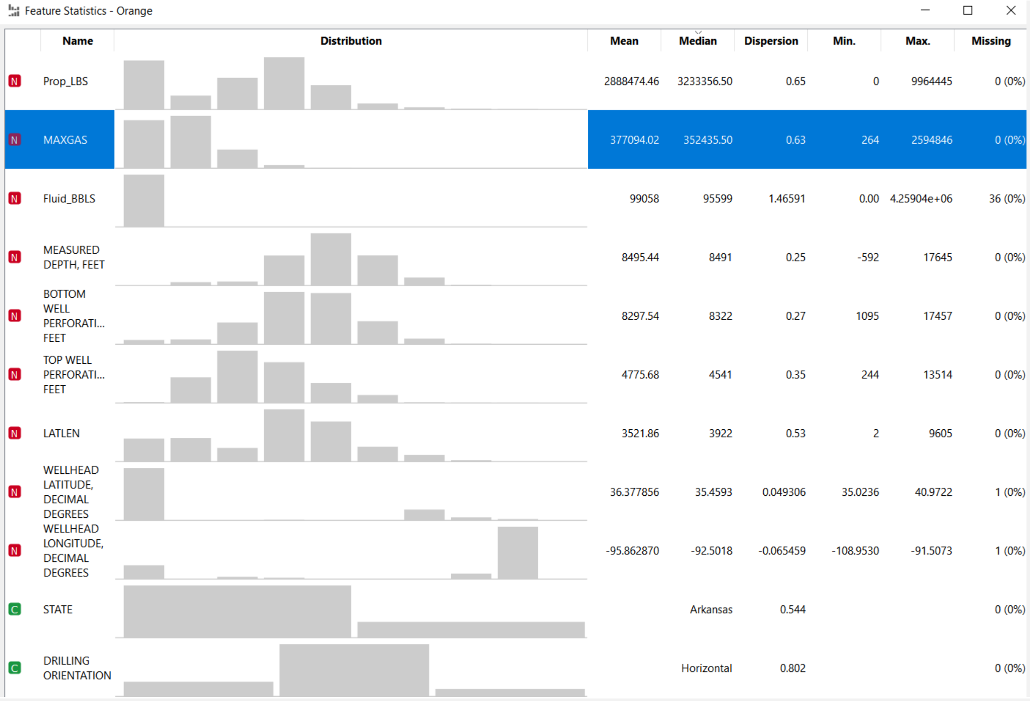
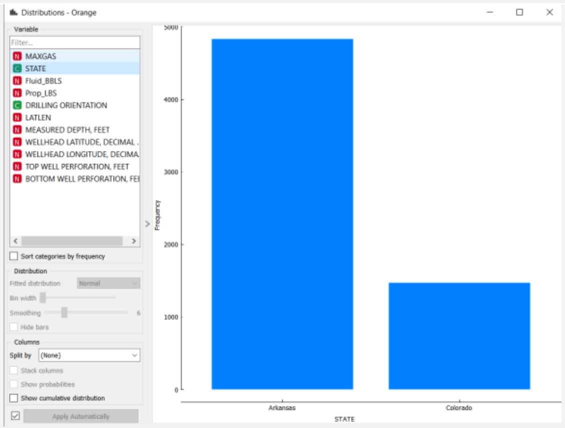
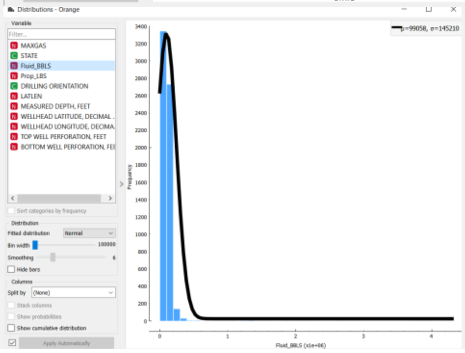


2. Inspect the **File** widget and complete the following table.

Total instances	6310
Dimensionality of the data set	 Rows: 6310 Columns: 14
Predictors (features)	10

Target variable	MAXGAS

3. Use the **Data Table, Feature Statistics, Distribution, Scatter Plot & Correlations and Scatter Plot** widgets to complete the table below:

Min and Max of predictor variables	
Do you notice any data inconsistency?	Yes, like depth -592 (depth cannot be negative, below depth)
Distribution of the predictor variables:	 

Feature Statistics - Orange

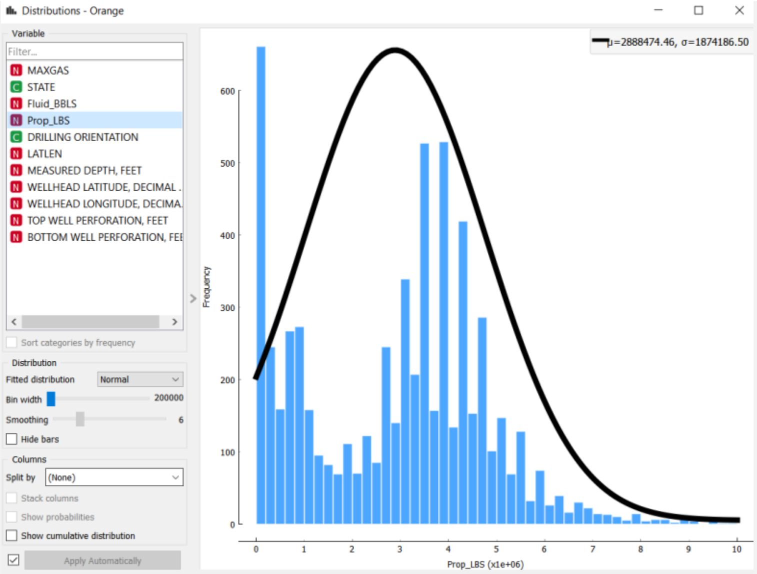
	Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
N	Prop_LBS		2888474.46	3233356.50	0.65	0	9964445	0 (0%)
N	MAXGAS		377094.02	352435.50	0.63	264	2594846	0 (0%)
N	Fluid_BBLS		99058	95599	1.46591	0.00	4.25904e+06	36 (0%)
N	MEASURED DEPTH, FEET		8495.44	8491	0.25	-592	17645	0 (0%)
N	BOTTOM WELL PERFORATI... FEET		8297.54	8322	0.27	1095	17457	0 (0%)
N	TOP WELL PERFORATI... FEET		4775.68	4541	0.35	244	13514	0 (0%)
N	LATLEN		3521.86	3922	0.53	2	9605	0 (0%)
N	WELLHEAD LATITUDE, DECIMAL DEGREES		36.377856	35.4593	0.049306	35.0236	40.9722	1 (0%)
N	WELLHEAD LONGITUDE, DECIMAL DEGREES		-95.862870	-92.5018	-0.065459	-108.9530	-91.5073	1 (0%)
C	STATE			Arkansas	0.544			0 (0%)
C	DRILLING ORIENTATION			Horizontal	0.802			0 (0%)

Distributions - Orange

□ Sort categories by frequency
Fitted distribution: Normal
Bin width
Smoothing: 6
□ Hide bars
Split by: (None)
□ Stack columns
□ Show probabilities
□ Show cumulative distribution
☑ Apply Automatically

Distributions - Orange

□ Sort categories by frequency
Fitted distribution: Normal
Bin width: 100000
Smoothing: 6
□ Hide bars
Split by: (None)
□ Stack columns
□ Show probabilities
□ Show cumulative distribution
☑ Apply Automatically



Distributions - Orange

Variable

Filter...

- N MAXGAS
- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL .
- N WELLHEAD LONGITUDE, DECIMA
- N TOP WELL PERFORATION, FEET
- N BOTTOM WELL PERFORATION, FEE

☐ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width: 200000

Smoothing: 6

☐ Hide bars

Columns

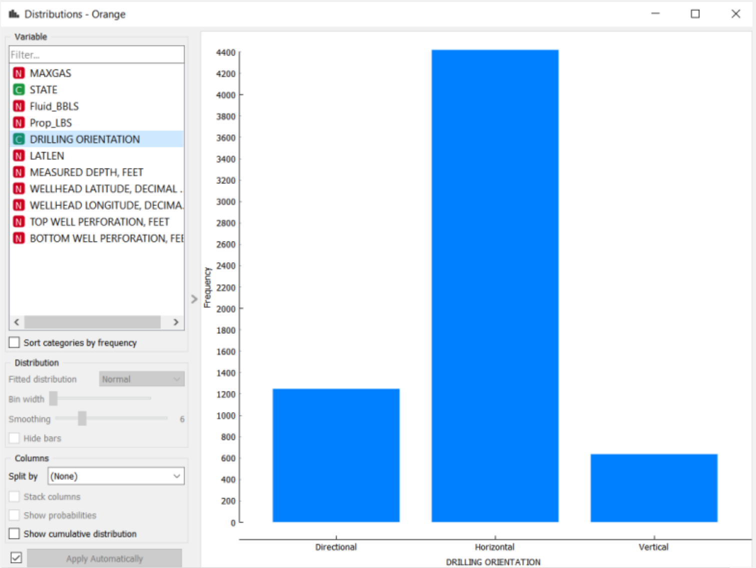
Split by: (None)

☐ Stack columns

☐ Show probabilities

☐ Show cumulative distribution

☑ Apply Automatically



Distributions - Orange

Variable

Filter...

- N MAXGAS
- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL .
- N WELLHEAD LONGITUDE, DECIMA
- N TOP WELL PERFORATION, FEET
- N BOTTOM WELL PERFORATION, FEE

☐ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width

Smoothing: 6

☐ Hide bars

Columns

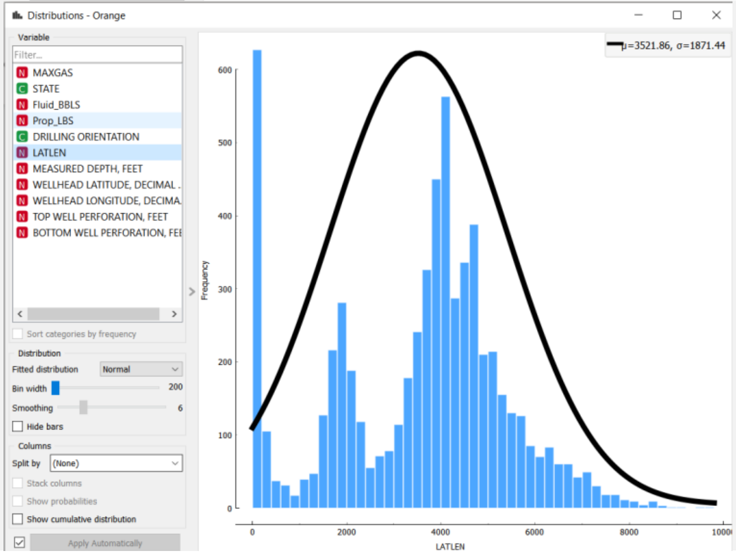
Split by: (None)

☐ Stack columns

☐ Show probabilities

☐ Show cumulative distribution

☑ Apply Automatically



Distributions - Orange

Variable

Filter...

- N MAXGAS
- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL .
- N WELLHEAD LONGITUDE, DECIMA
- N TOP WELL PERFORATION, FEET
- N BOTTOM WELL PERFORATION, FEE

☐ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width: 200

Smoothing: 6

☐ Hide bars

Columns

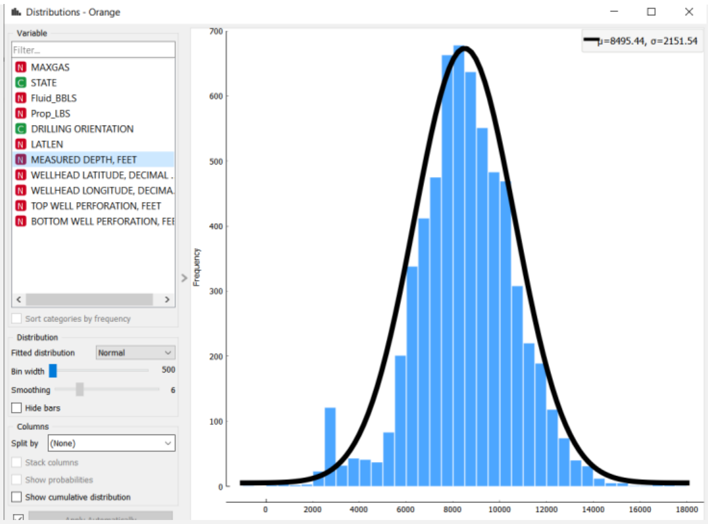
Split by: (None)

☐ Stack columns

☐ Show probabilities

☐ Show cumulative distribution

☑ Apply Automatically



Fitted distribution: Normal

Bin width: 500

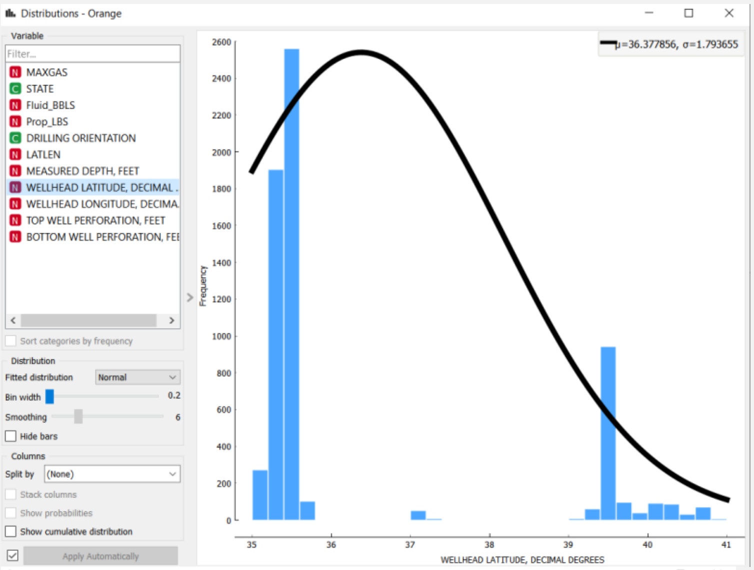
Smoothing: 6

□ Hide bars

Split by: (None)

□ Show cumulative distribution

μ=8495.44, σ=2151.54



Fitted distribution: Normal

Bin width: 0.2

Smoothing: 6

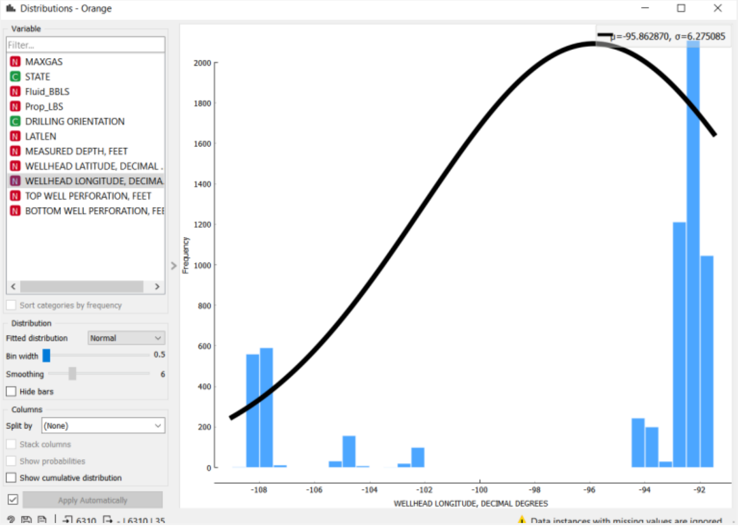
□ Hide bars

Split by: (None)

□ Show cumulative distribution

☑ Apply Automatically

μ=36.377856, σ=1.793655



Fitted distribution: Normal

Bin width: 0.5

Smoothing: 6

□ Hide bars

Split by: (None)

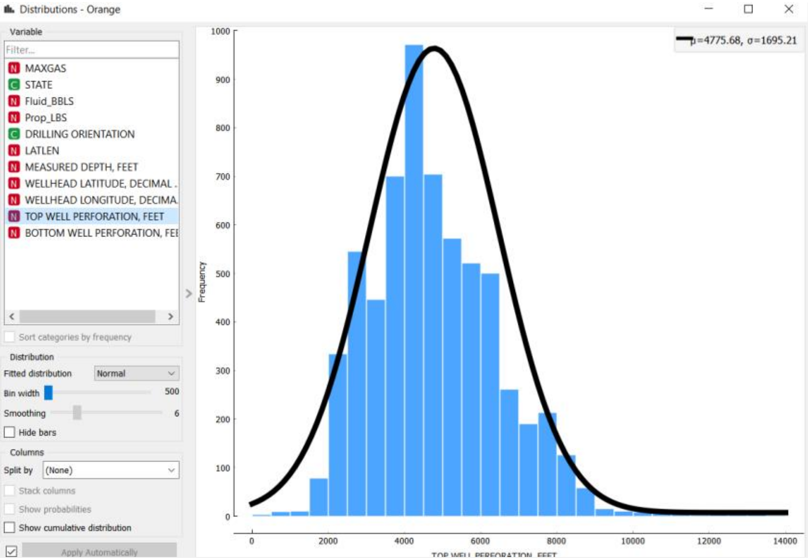
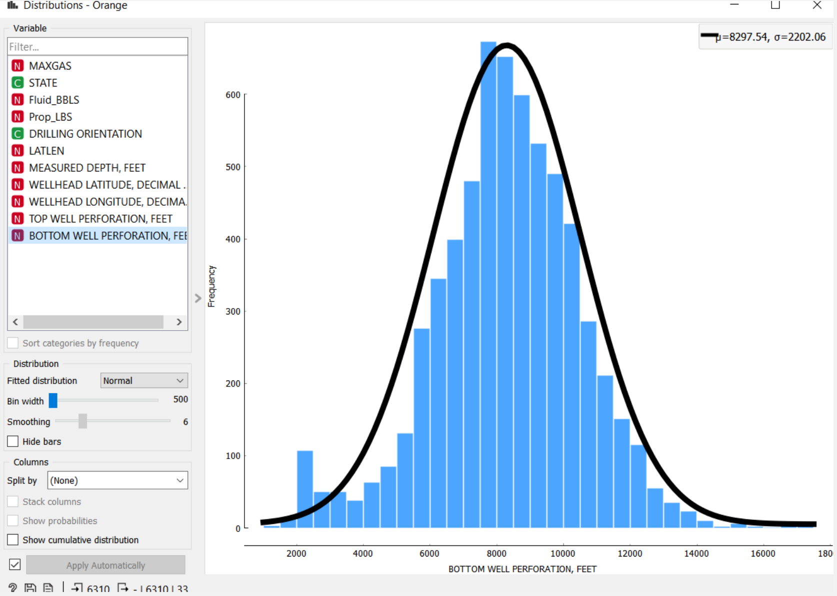
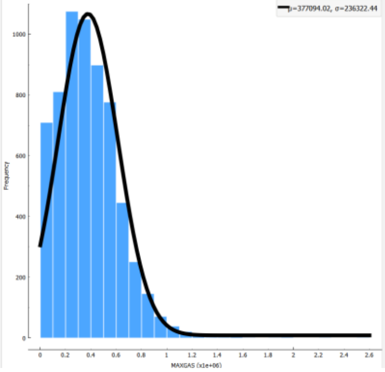
□ Show cumulative distribution

☑ Apply Automatically

μ=-95.862870, σ=6.275085

6310 | - | 6310 | 35

Data instances with missing values are ignored

	
	
Min, Max and distribution of the target variable, Max_Gas	Min:264, Max: 2594846, skewed distribution 

Distributions - Orange

Variable

Filter...

- N MAXGAS
- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL .
- N WELLHEAD LONGITUDE, DECIMA
- N TOP WELL PERFORATION, FEET
- N BOTTOM WELL PERFORATION, FEI

□ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width: 500

Smoothing: 6

□ Hide bars

Columns

Split by: (None)

□ Stack columns

□ Show probabilities

□ Show cumulative distribution

☑ Apply Automatically

μ=4775.68, σ=1695.21

Distributions - Orange

Variable

Filter...

- N MAXGAS
- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL .
- N WELLHEAD LONGITUDE, DECIMA
- N TOP WELL PERFORATION, FEET
- N BOTTOM WELL PERFORATION, FEI

□ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width: 500

Smoothing: 6

□ Hide bars

Columns

Split by: (None)

□ Stack columns

□ Show probabilities

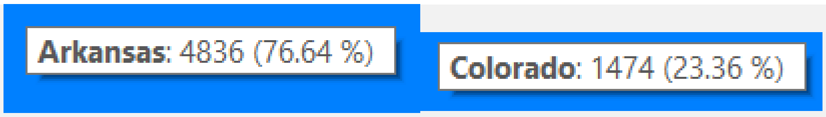
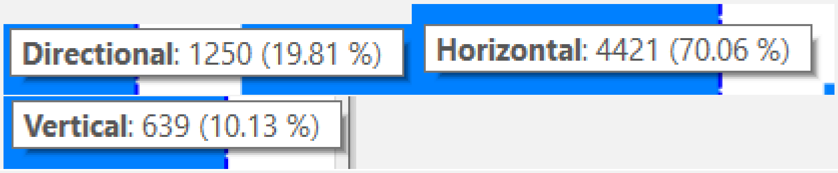
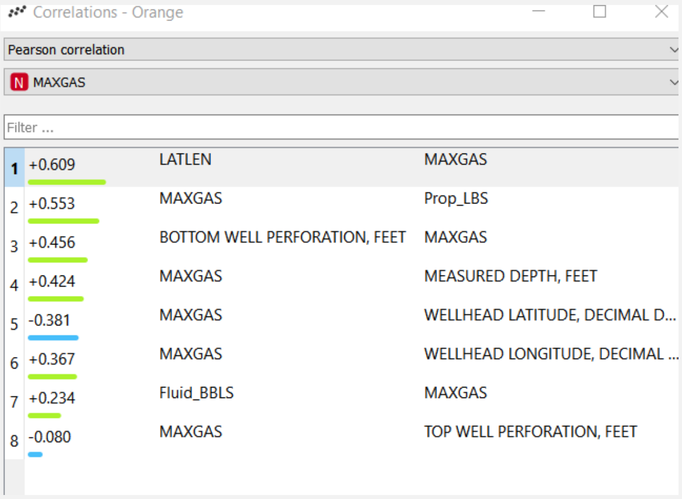
□ Show cumulative distribution

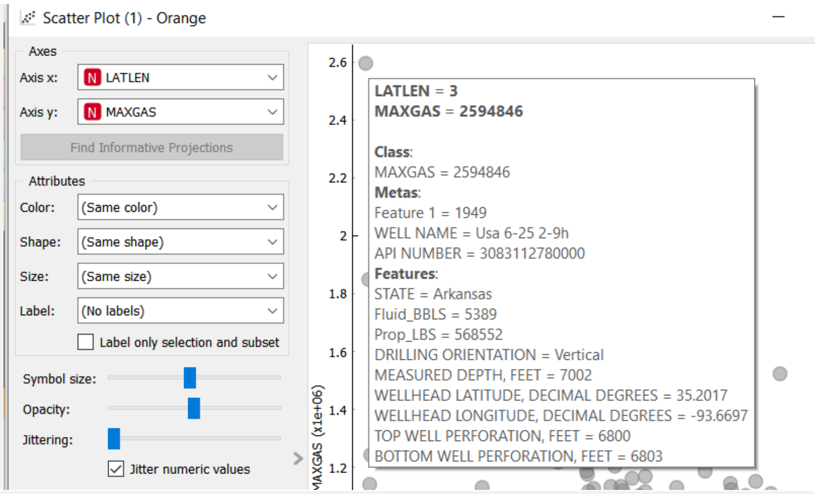
☑ Apply Automatically

6310 | 6310 | 33

μ=8297.54, σ=2202.06

μ=377094.02, σ=236322.44

% of wells from the two states (shale plays) and drilling orientation	  Directional: 1250 (19.81 %) Horizontal: 4421 (70.06 %) Vertical: 639 (10.13 %)
Any missing values?	yes
From the **Scatter Plot**, what are the predictors that you believe would have the highest impact on predicting Max_Gas?	LATLEN (60% strongest correlation), Prop_LBS, BOTTOM WELL…  Correlations - Orange Pearson correlation N MAXGAS Filter ... 1 +0.609 LATLEN MAXGAS 2 +0.553 MAXGAS Prop_LBS 3 +0.456 BOTTOM WELL PERFORATION, FEET MAXGAS 4 +0.424 MAXGAS MEASURED DEPTH, FEET 5 -0.381 MAXGAS WELLHEAD LATITUDE, DECIMAL D... 6 +0.367 MAXGAS WELLHEAD LONGITUDE, DECIMAL ... 7 +0.234 Fluid_BBLS MAXGAS 8 -0.080 MAXGAS TOP WELL PERFORATION, FEET
How would you describe the relationships for these predictors with the target? Do you notice outliers in the data?	Increase of predictors usually mean more gas productions (MAXGAS) Yes, for example: X: LATLEN, Y: MAXGAS, and has an outlier close to 2.6



Scatter Plot (1) - Orange

Axes

Axis x: N LATLEN

Axis y: N MAXGAS

Find Informative Projections

Attributes

Color: (Same color)

Shape: (Same shape)

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

Symbol size:

Opacity:

Jittering:

☑ Jitter numeric values

LATLEN = **3**
MAXGAS = **2594846**

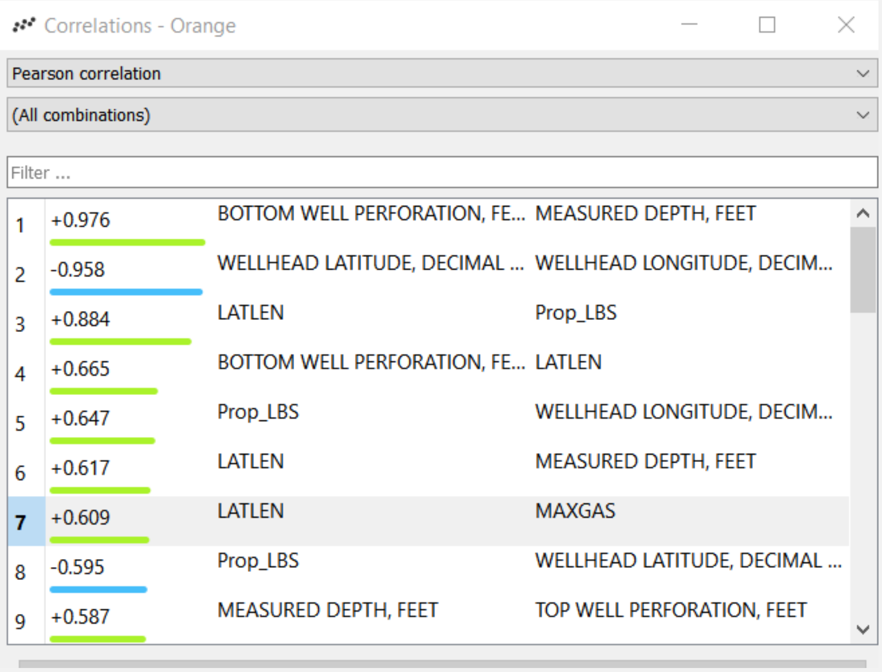
Class:
MAXGAS = 2594846
Metas:
Feature 1 = 1949
WELL NAME = Usa 6-25 2-9h
API NUMBER = 3083112780000
Features:
STATE = Arkansas
Fluid_BBLS = 5389
Prop_LBS = 568552
DRILLING ORIENTATION = Vertical
MEASURED DEPTH, FEET = 7002
WELLHEAD LATITUDE, DECIMAL DEGREES = 35.2017
WELLHEAD LONGITUDE, DECIMAL DEGREES = -93.6697
TOP WELL PERFORATION, FEET = 6800
BOTTOM WELL PERFORATION, FEET = 6803

From the **correlations** widget, what do you think about the correlation between predictors?

LATLEN 60% strongest correlation with MAXGAS, second is Prop_LBS (55%)

some have really strong correlations with each other (LATLEn and Prop_LBS 88%)

Each predictors have their strong correlations, expected to be independent



Correlations - Orange

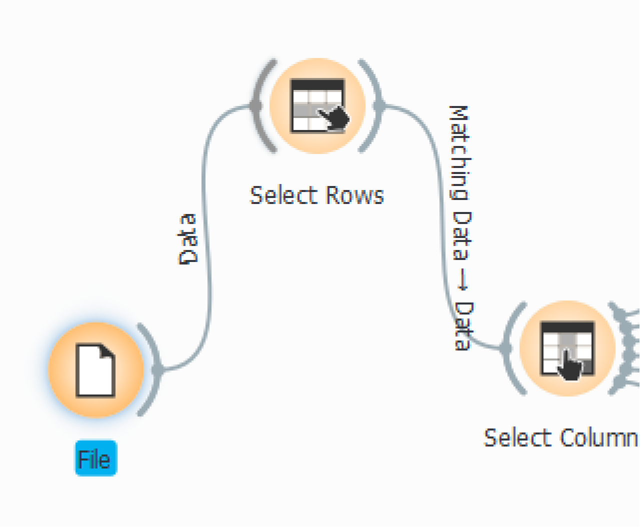
Pearson correlation

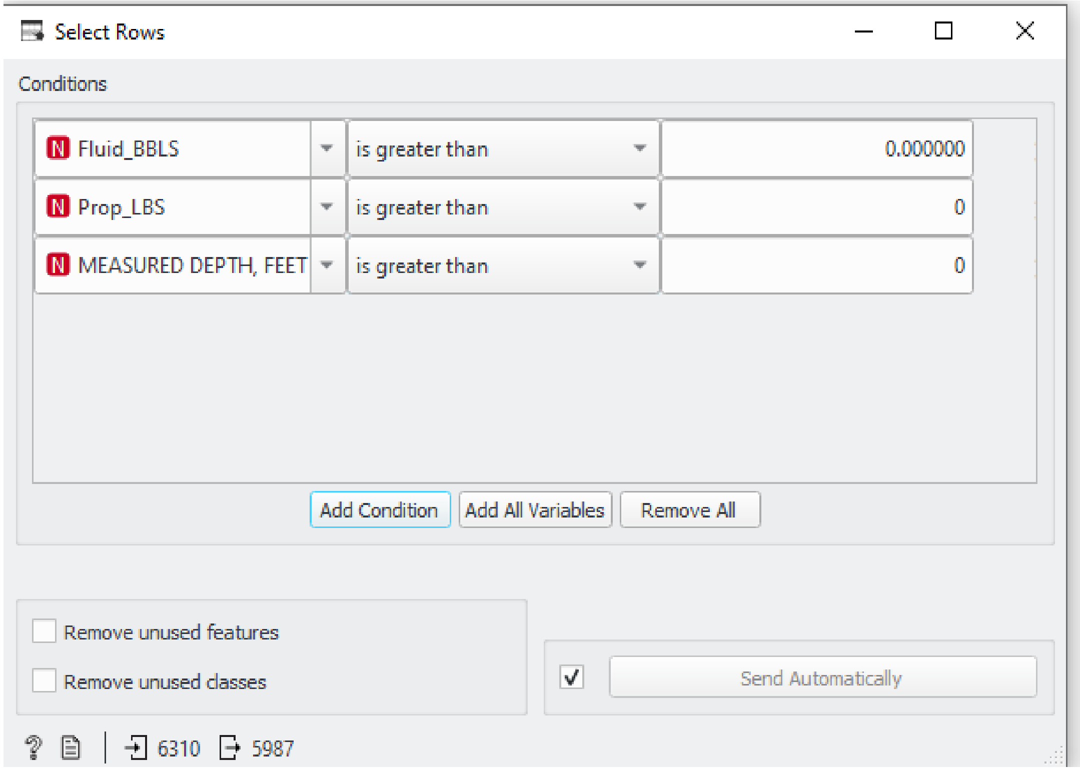
(All combinations)

Filter ...

1	+0.976	BOTTOM WELL PERFORATION, FE...	MEASURED DEPTH, FEET
2	-0.958	WELLHEAD LATITUDE, DECIMAL ...	WELLHEAD LONGITUDE, DECIM...
3	+0.884	LATLEN	Prop_LBS
4	+0.665	BOTTOM WELL PERFORATION, FE...	LATLEN
5	+0.647	Prop_LBS	WELLHEAD LONGITUDE, DECIM...
6	+0.617	LATLEN	MEASURED DEPTH, FEET
7	+0.609	LATLEN	MAXGAS
8	-0.595	Prop_LBS	WELLHEAD LATITUDE, DECIMAL ...
9	+0.587	MEASURED DEPTH, FEET	TOP WELL PERFORATION, FEET

4. Address data inconsistency by adding **Select Rows** widget and filtering them out.




Select Rows

Conditions

N Fluid_BBLS	is greater than	0.000000
N Prop_LBS	is greater than	0
N MEASURED DEPTH, FEET	is greater than	0

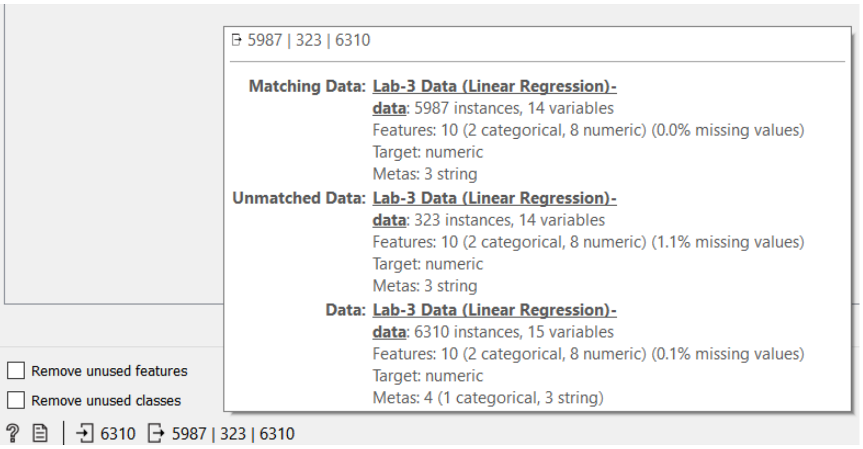
□ Remove unused features

□ Remove unused classes

☑ Send Automatically

6310 5987

How many instances were filtered out in the above step? Eliminated 323 rows



5987 | 323 | 6310

Matching Data: Lab-3 Data (Linear Regression)-data: 5987 instances, 14 variables
Features: 10 (2 categorical, 8 numeric) (0.0% missing values)
Target: numeric
Metas: 3 string

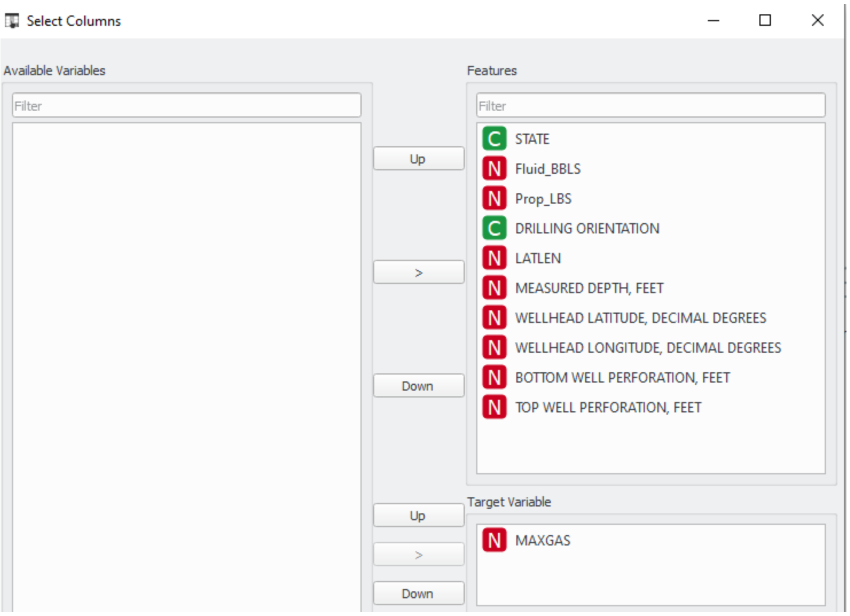
Unmatched Data: Lab-3 Data (Linear Regression)-data: 323 instances, 14 variables
Features: 10 (2 categorical, 8 numeric) (1.1% missing values)
Target: numeric
Metas: 3 string

Data: Lab-3 Data (Linear Regression)-data: 6310 instances, 15 variables
Features: 10 (2 categorical, 8 numeric) (0.1% missing values)
Target: numeric
Metas: 4 (1 categorical, 3 string)

□ Remove unused features
□ Remove unused classes

6310 | 5987 | 323 | 6310

5. Open **Select Columns** widget and confirm selection as below:



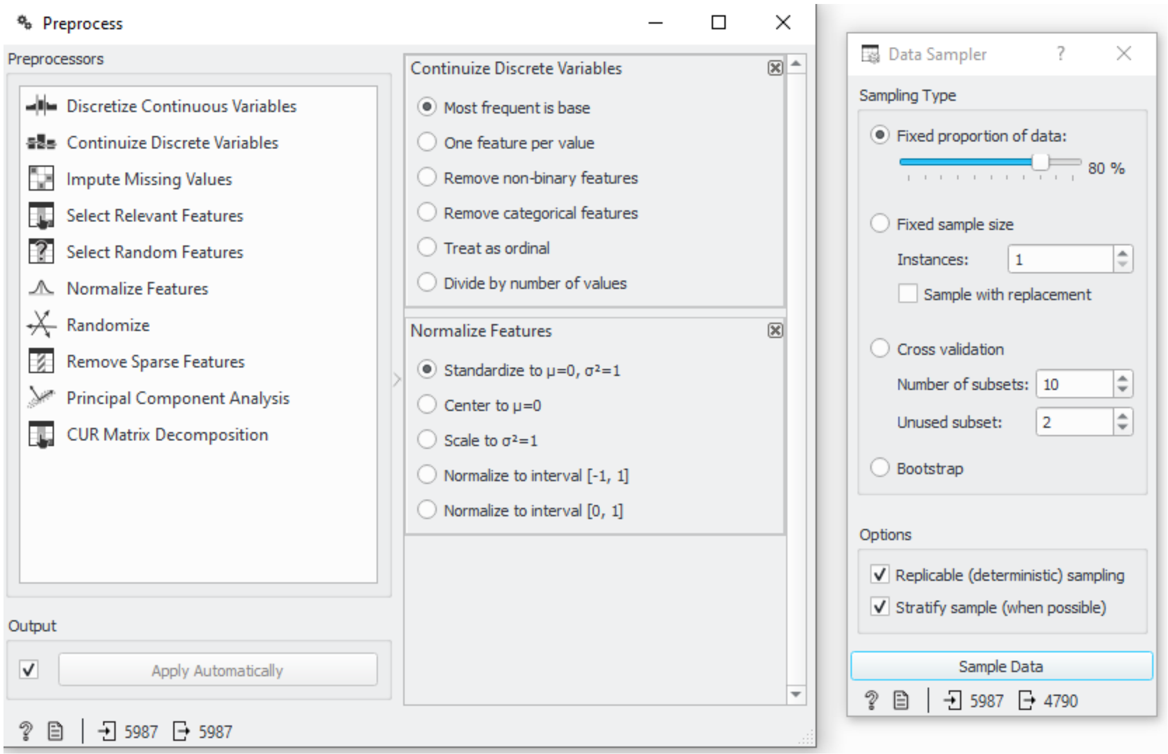
Select Columns

Available Variables

Features
- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL DEGREES
- N WELLHEAD LONGITUDE, DECIMAL DEGREES
- N BOTTOM WELL PERFORATION, FEET
- N TOP WELL PERFORATION, FEET

Target Variable
- N MAXGAS

6. Inspect the **Preprocess** and **Data Sampler** widgets.



Preprocess

Preprocessors

- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Remove Sparse Features
- Principal Component Analysis
- CUR Matrix Decomposition

Continuize Discrete Variables

● Most frequent is base
○ One feature per value
○ Remove non-binary features
○ Remove categorical features
○ Treat as ordinal
○ Divide by number of values

Normalize Features

● Standardize to μ=0, σ²=1
○ Center to μ=0
○ Scale to σ²=1
○ Normalize to interval [-1, 1]
○ Normalize to interval [0, 1]

Output

☑ Apply Automatically

5987 5987

Data Sampler

Sampling Type

● Fixed proportion of data: 80 %
○ Fixed sample size
Instances: 1
□ Sample with replacement
○ Cross validation
Number of subsets: 10
Unused subset: 2
○ Bootstrap

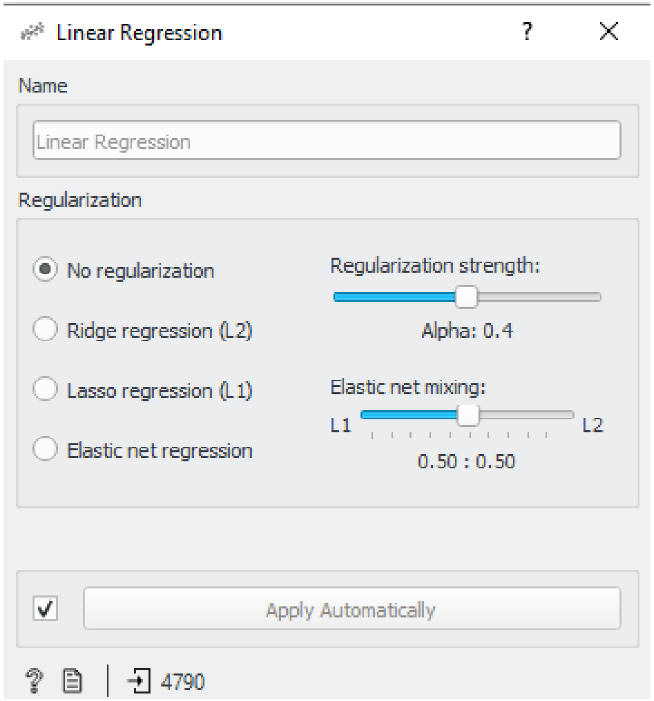
Options

☑ Replicable (deterministic) sampling
☑ Stratify sample (when possible)

Sample Data

5987 4790

7. Open the **Linear Regression** widget and inspect the same:



Linear Regression

Name

Linear Regression

Regularization

● No regularization
○ Ridge regression (L2)
○ Lasso regression (L1)
○ Elastic net regression

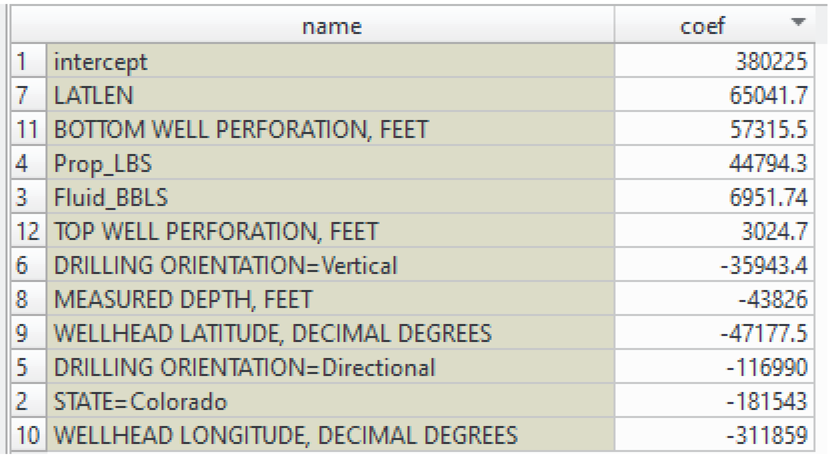
Regularization strength:
Alpha: 0.4

Elastic net mixing:
L1 L2
0.50 : 0.50

☑ Apply Automatically

4790

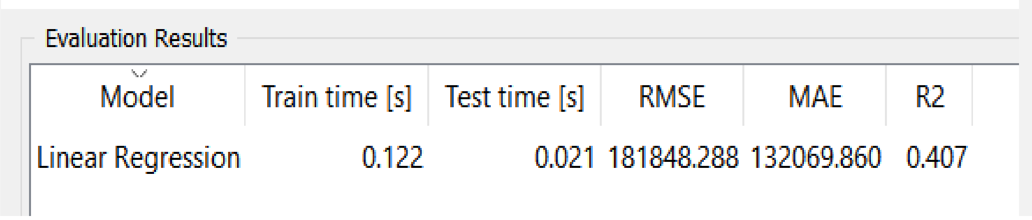
8. Open **Data Table** widget connected to the **Linear Regression** widget and inspect the parameter coefficients.



	name	coef
1	intercept	380225
7	LATLEN	65041.7
11	BOTTOM WELL PERFORATION, FEET	57315.5
4	Prop_LBS	44794.3
3	Fluid_BBLS	6951.74
12	TOP WELL PERFORATION, FEET	3024.7
6	DRILLING ORIENTATION=Vertical	-35943.4
8	MEASURED DEPTH, FEET	-43826
9	WELLHEAD LATITUDE, DECIMAL DEGREES	-47177.5
5	DRILLING ORIENTATION=Directional	-116990
2	STATE=Colorado	-181543
10	WELLHEAD LONGITUDE, DECIMAL DEGREES	-311859

What can you say about the relationships between the different predictors and the response variable (Max_Gas)?	They influence or relate some way (LATLEN most important variable and strongest influence); vertical wells will have a degraded performance while horizontal perform better; measured depth is actually having negative influence to the production

9. Open **Test and Score** widget and complete the table below:

What are the evaluation metrics you see?	

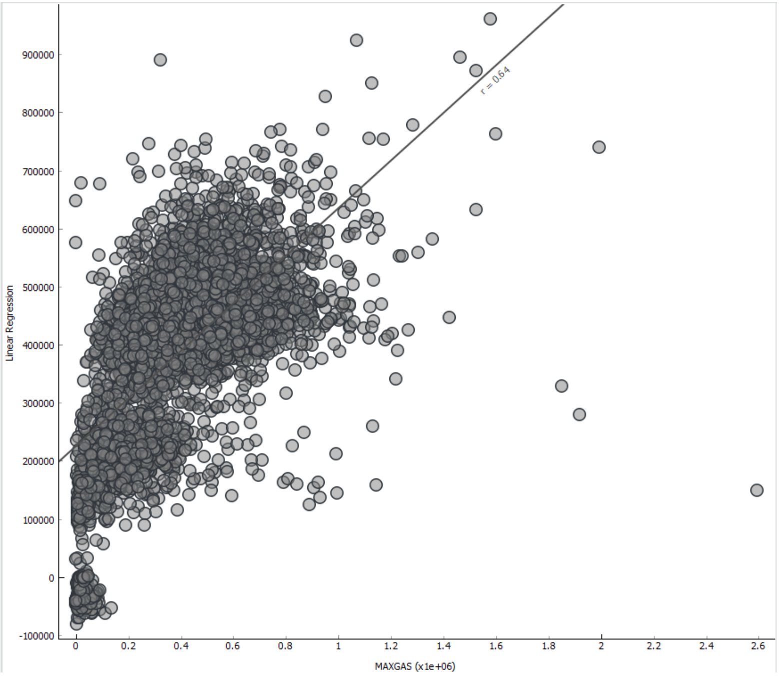
Evaluation Results

Model	Train time [s]	Test time [s]	RMSE	MAE	R2
Linear Regression	0.122	0.021	181848.288	132069.860	0.407

Model	RMSE	MAE	R2
Linear Regression (cross validation , 10 folds	181848.288	132069.860	0.407
Linear Regression (Test on test data)	182165.880	131290.348	0.396

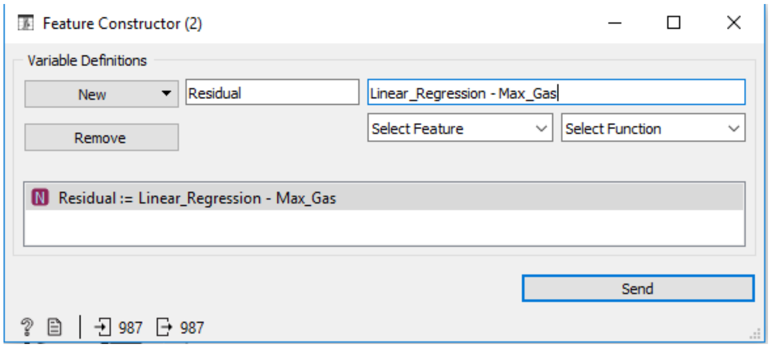
How would you interpret this model based on R2 and cross validation?	Decreased percentage, and given R2 is 40 percent vs 100 (or high number), it is not a perfect model

Open the **Scatter Plot** widget connected to **Predictions widget.** Select Max_Gas as X-axis and Linear Regression as Y-axis.



What do you think about the prediction quality?	Not too perfect since not all the plots are lined to the r and likely not deploy the model, but still wouldn't discard this to see what can be improved

10. As part of model diagnostics in linear regression, it is standard practice to visualize the residual (error) plot and look for patterns to understand model validity.
Residual variable is constructed as shown below using the **Feature Constructor** widget:

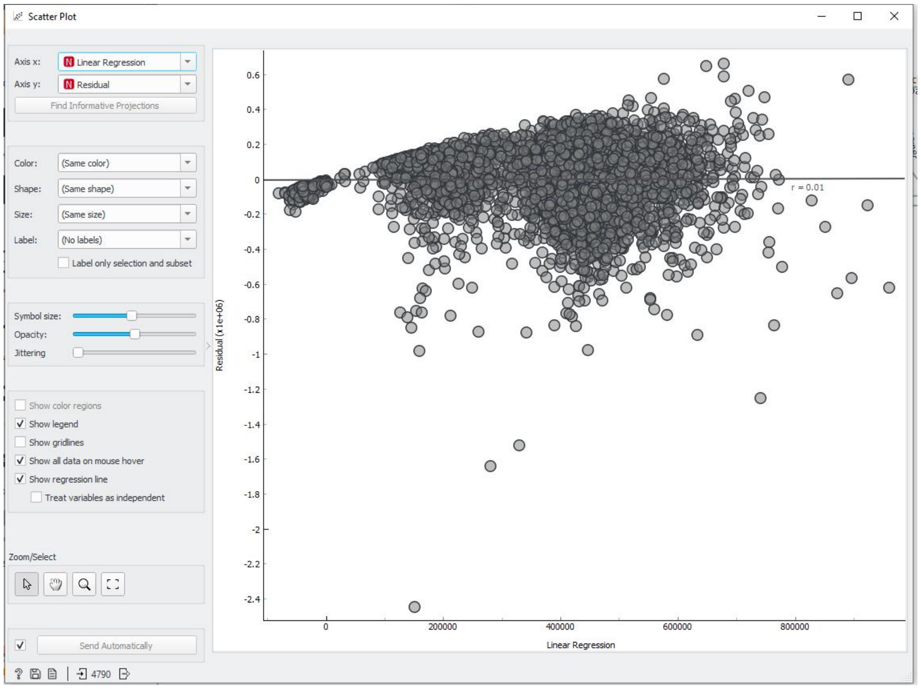


Residual

Linear_Regression - Max_Gas

N Residual := Linear_Regression - Max_Gas

11. Open **Scatter Plot** widget and select Linear Regression (X axis)-(Predicted Max_Gas from the regression model) and Residual (Y axis). Plot would look as below:



Axis x: Linear Regression
Axis y: Residual

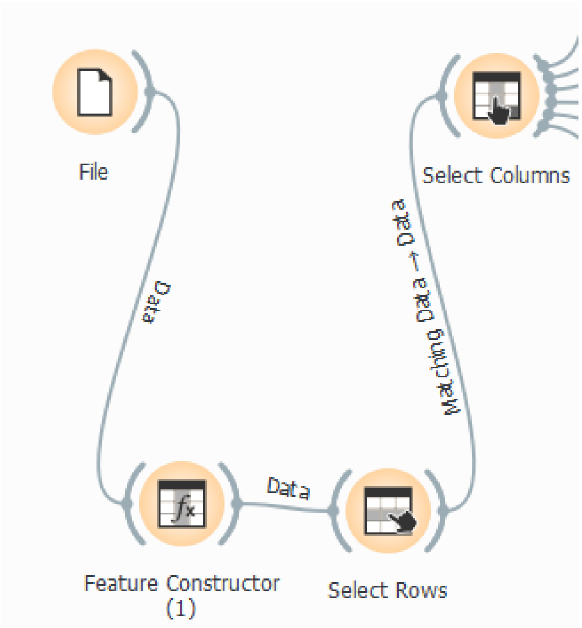
Color: (Same color)
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
☐ Label only selection and subset

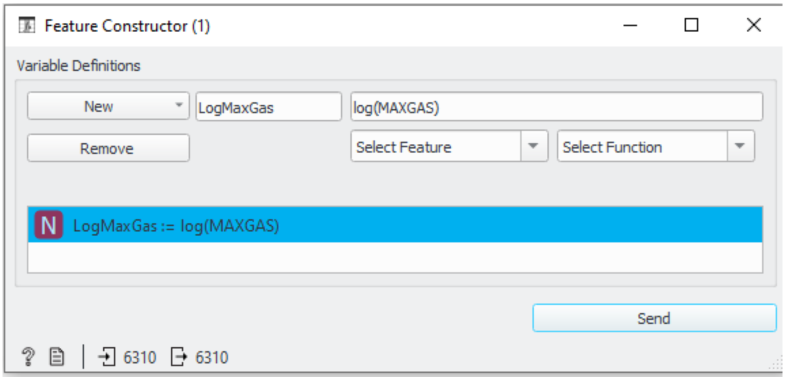
☐ Show color regions
☑ Show legend
☐ Show gridlines
☑ Show all data on mouse hover
☑ Show regression line
☐ Treat variables as independent

☑ Send Automatically

Do you see a pattern? What does this indicate? What can you do to address this issue?	Center line running through 0, thus a corn/megaphone pattern Average out to mean; when you train a linear regression model, for it to be valid, the error should have a normal distribution with a mean of 0 Indicates that tho it looks okay, it is not quite valid Address by apply transformation (look at log or square root of variable) to your target variable (changes distribution of target to make it valid more likely)

12. Let us transform the target variable Max_Gas using Log. Add another **Feature Constructor** widget near the **File widget.** Construct Log transformation as shown below:





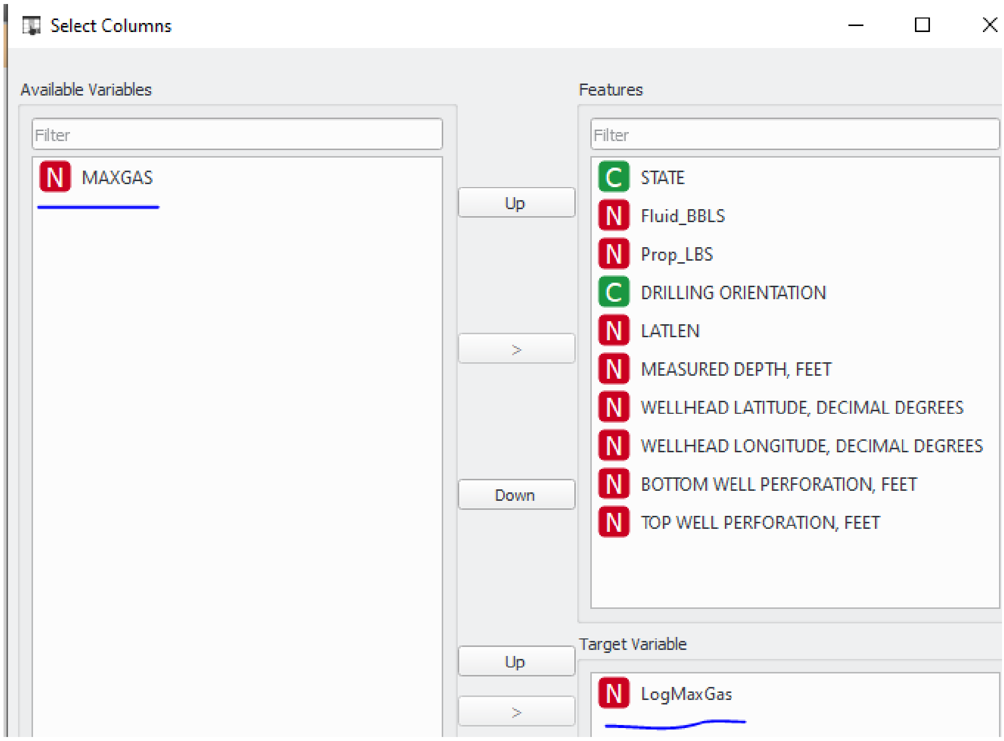
LogMaxGas

log(MAXGAS)

N LogMaxGas := log(MAXGAS)

6310 6310

Modify **Select Columns** widget as below:



Select Columns

Available Variables

Filter

- N MAXGAS

Features

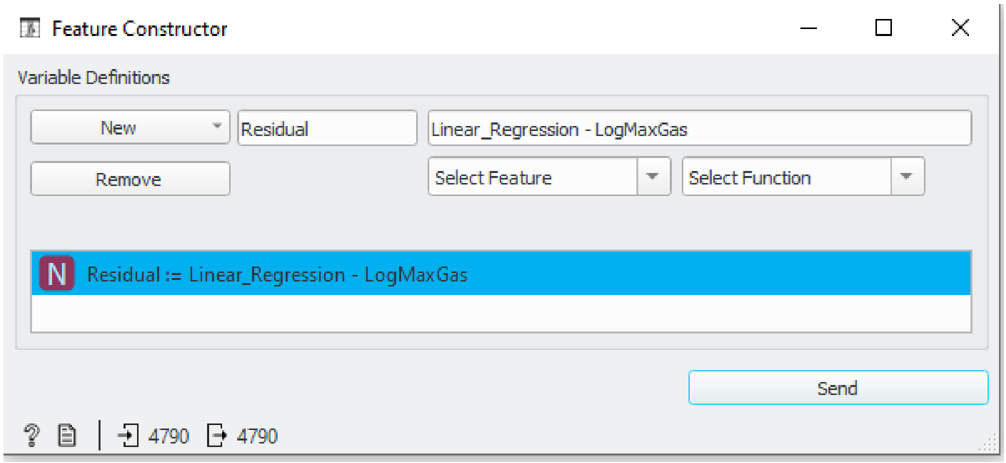
Filter

- C STATE
- N Fluid_BBLS
- N Prop_LBS
- C DRILLING ORIENTATION
- N LATLEN
- N MEASURED DEPTH, FEET
- N WELLHEAD LATITUDE, DECIMAL DEGREES
- N WELLHEAD LONGITUDE, DECIMAL DEGREES
- N BOTTOM WELL PERFORATION, FEET
- N TOP WELL PERFORATION, FEET

Target Variable

- N LogMaxGas

Update the second Feature constructor (farthest right) to calculate the residual correctly:



Feature Constructor

Variable Definitions

Residual

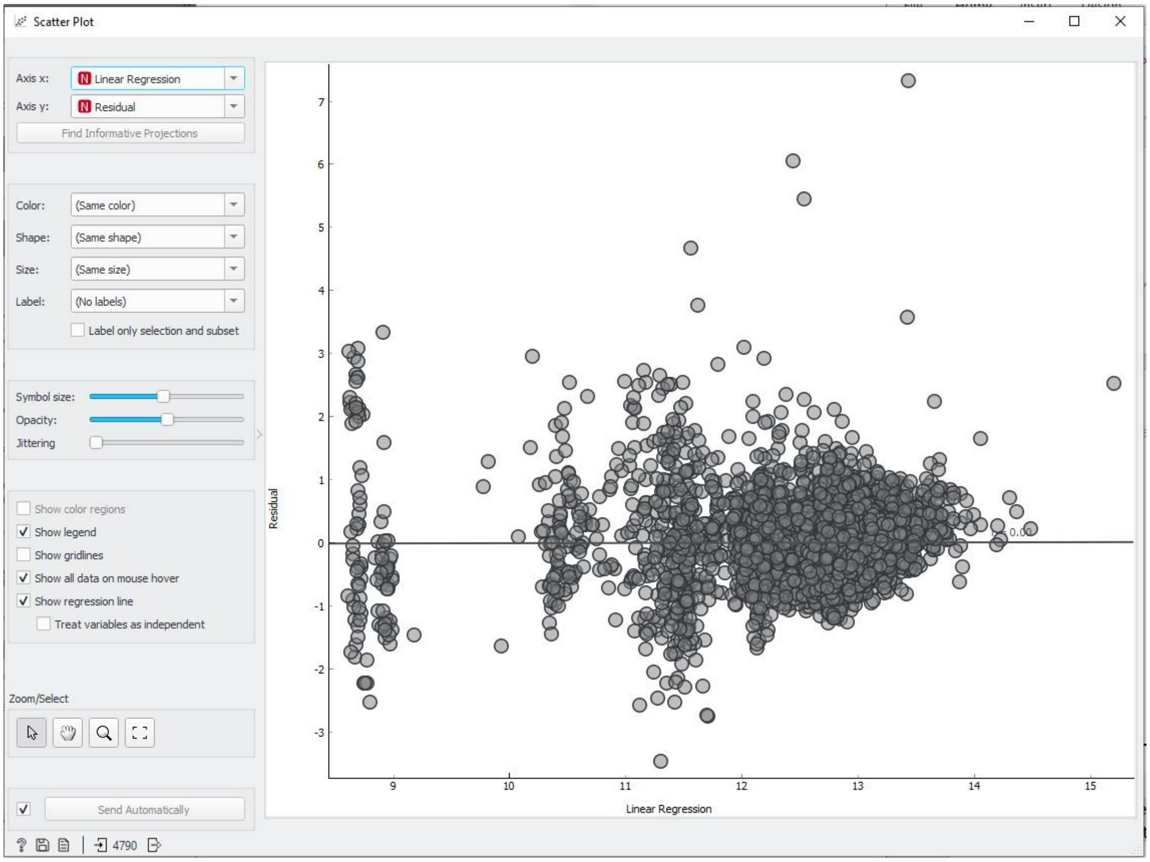
Linear_Regression - LogMaxGas

Select Feature

Select Function

N Residual := Linear_Regression - LogMaxGas

Open and visualize the residual plot to see the effect of the Log transformation of Max_Gas target variable:



Axis x: Linear Regression
Axis y: Residual

Color: (Same color)
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
□ Label only selection and subset

Symbol size:
Opacity:
Jittering

□ Show color regions
☑ Show legend
□ Show gridlines
☑ Show all data on mouse hover
☑ Show regression line
□ Treat variables as independent

☑ Send Automatically

4790

13. Open **Test and Score** widget and complete the table below. Notice the improvement of the model performance.

Model	RMSE	MAE	R2
Linear Regression, LogMaxGas (Cross Validation, 10 folds)	0.590	0.428	0.655

14. If you do a prediction using this model, what is an important consideration before using the predicted results?

RMS value not the same as before because now a log-transformed variable, thus cannot compare the errors the same way (numerically not the same); but the R2 (r-square) should give an indication that the model improved significantly