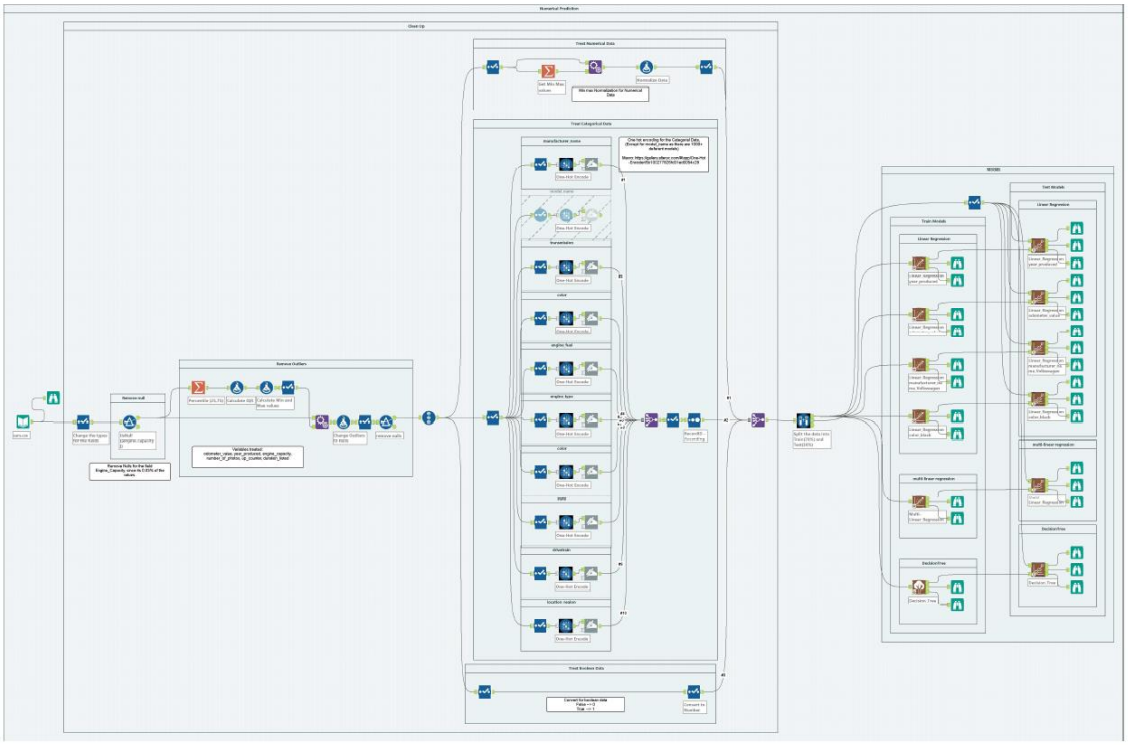
M6 Practical Challenge: Predictive Modeling with Alteryx

Table of Contents

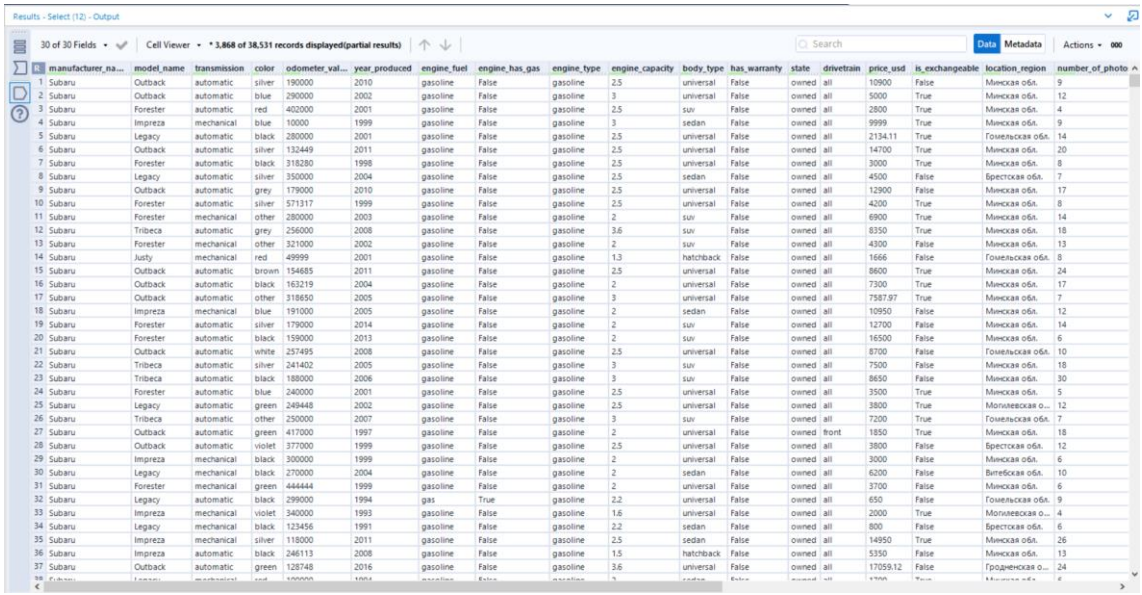
Alejandro C Parra Garcia

Numerical Prediction

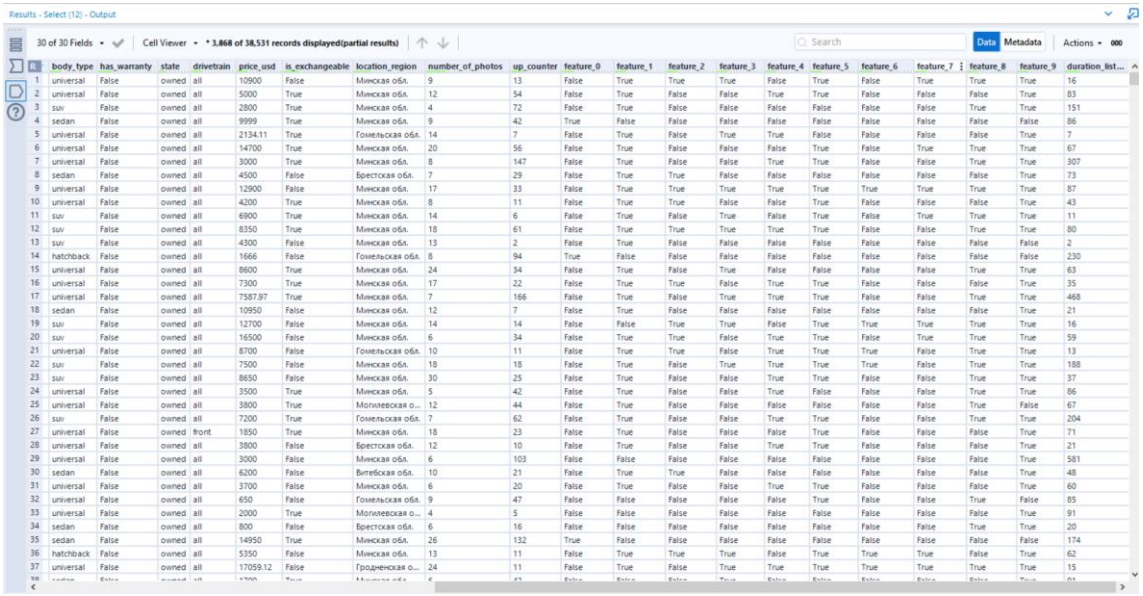


Imported Data:

The imported values are:



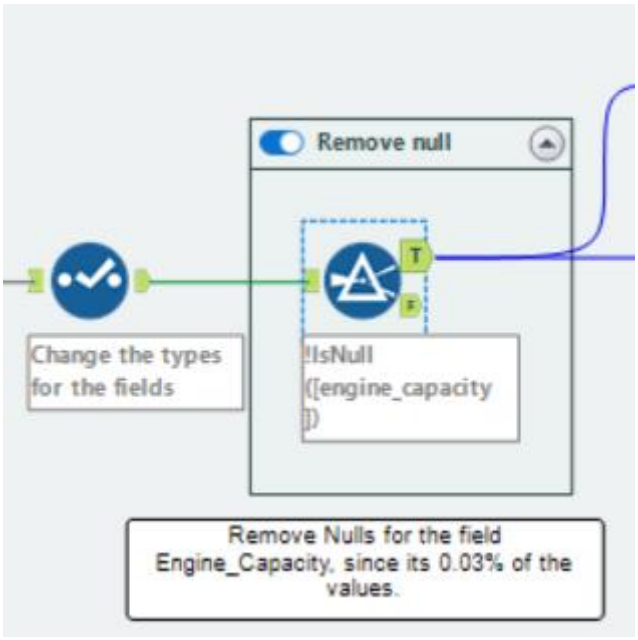
	manufacturer_na...	model_name	transmission	color	odometer_val...	year_produced	engine_fuel	engine_has_gas	engine_type	engine_capacity	body_type	has_warranty	state	drivetrain	price_usd	is_exchangeable	location_region	number_of_photo
1	Subaru	Outback	automatic	silver	190000	2010	gasoline	False	gasoline	2.5	universal	False	owned	all	10900	True	Минская обл.	9
2	Subaru	Outback	automatic	blue	290000	2002	gasoline	False	gasoline	3	universal	False	owned	all	5000	True	Минская обл.	12
3	Subaru	Forester	automatic	red	402000	2001	gasoline	False	gasoline	2.5	suv	False	owned	all	2800	True	Минская обл.	4
4	Subaru	Impreza	mechanical	blue	10000	1999	gasoline	False	gasoline	3	sedan	False	owned	all	9999	True	Минская обл.	9
5	Subaru	Legacy	automatic	black	280000	2001	gasoline	False	gasoline	2.5	universal	False	owned	all	2134.11	True	Гомельская обл.	14
6	Subaru	Outback	automatic	silver	132449	2011	gasoline	False	gasoline	2.5	universal	False	owned	all	14700	True	Минская обл.	20
7	Subaru	Forester	automatic	black	318280	1998	gasoline	False	gasoline	2.5	universal	False	owned	all	3000	True	Минская обл.	8
8	Subaru	Legacy	automatic	silver	350000	2004	gasoline	False	gasoline	2.5	sedan	False	owned	all	4500	False	Брестская обл.	7
9	Subaru	Outback	automatic	grey	179000	2010	gasoline	False	gasoline	2.5	universal	False	owned	all	12900	False	Минская обл.	17
10	Subaru	Forester	automatic	silver	571317	1999	gasoline	False	gasoline	2.5	universal	False	owned	all	4200	True	Минская обл.	8
11	Subaru	Forester	mechanical	other	280000	2003	gasoline	False	gasoline	2	suv	False	owned	all	6900	True	Минская обл.	14
12	Subaru	Tribeca	automatic	grey	256000	2008	gasoline	False	gasoline	3.6	suv	False	owned	all	8350	True	Минская обл.	18
13	Subaru	Forester	mechanical	other	321000	2002	gasoline	False	gasoline	2	suv	False	owned	all	4300	False	Минская обл.	13
14	Subaru	Justy	mechanical	red	49999	2001	gasoline	False	gasoline	1.3	hatchback	False	owned	all	1666	False	Гомельская обл.	8
15	Subaru	Outback	automatic	brown	154685	2011	gasoline	False	gasoline	2.5	universal	False	owned	all	8600	True	Минская обл.	24
16	Subaru	Outback	automatic	black	163219	2004	gasoline	False	gasoline	2	universal	False	owned	all	7300	True	Минская обл.	17
17	Subaru	Outback	automatic	other	318650	2005	gasoline	False	gasoline	3	universal	False	owned	all	7587.97	True	Минская обл.	7
18	Subaru	Impreza	mechanical	blue	191000	2005	gasoline	False	gasoline	2	sedan	False	owned	all	10950	False	Минская обл.	12
19	Subaru	Forester	automatic	silver	179000	2014	gasoline	False	gasoline	2	suv	False	owned	all	12700	False	Минская обл.	14
20	Subaru	Forester	automatic	black	159000	2013	gasoline	False	gasoline	2	suv	False	owned	all	16500	False	Минская обл.	6
21	Subaru	Outback	automatic	white	257495	2008	gasoline	False	gasoline	2.5	universal	False	owned	all	8700	False	Гомельская обл.	10
22	Subaru	Tribeca	automatic	silver	241402	2005	gasoline	False	gasoline	3	suv	False	owned	all	7500	False	Минская обл.	18
23	Subaru	Tribeca	automatic	black	188000	2006	gasoline	False	gasoline	3	suv	False	owned	all	8650	False	Минская обл.	30
24	Subaru	Forester	automatic	blue	240000	2001	gasoline	False	gasoline	2.5	universal	False	owned	all	3500	True	Минская обл.	5
25	Subaru	Legacy	automatic	green	249448	2002	gasoline	False	gasoline	2.5	universal	False	owned	all	3800	True	Могилевская о...	12
26	Subaru	Tribeca	automatic	other	250000	2007	gasoline	False	gasoline	3	suv	False	owned	all	7200	True	Гомельская обл.	7
27	Subaru	Outback	automatic	green	417000	1997	gasoline	False	gasoline	2	universal	False	owned	front	1850	True	Минская обл.	18
28	Subaru	Outback	automatic	violet	377000	1999	gasoline	False	gasoline	2.5	universal	False	owned	all	3800	False	Брестская обл.	12
29	Subaru	Impreza	mechanical	black	300000	1999	gasoline	False	gasoline	2	universal	False	owned	all	3000	False	Минская обл.	6
30	Subaru	Legacy	mechanical	black	270000	2004	gasoline	False	gasoline	2	sedan	False	owned	all	6200	False	Витебская обл.	10
31	Subaru	Forester	mechanical	green	444444	1999	gasoline	False	gasoline	2	universal	False	owned	all	3700	False	Минская обл.	6
32	Subaru	Legacy	automatic	black	299000	1994	gas	True	gasoline	2.2	universal	False	owned	all	650	False	Гомельская обл.	9
33	Subaru	Impreza	mechanical	violet	340000	1993	gasoline	False	gasoline	1.6	universal	False	owned	all	2000	True	Могилевская о...	4
34	Subaru	Legacy	mechanical	black	123456	1991	gasoline	False	gasoline	2.2	sedan	False	owned	all	800	False	Брестская обл.	6
35	Subaru	Impreza	mechanical	silver	118000	2011	gasoline	False	gasoline	2.5	sedan	False	owned	all	14950	True	Минская обл.	26
36	Subaru	Impreza	automatic	black	246113	2008	gasoline	False	gasoline	1.5	hatchback	False	owned	all	5350	False	Минская обл.	13
37	Subaru	Outback	automatic	green	128748	2016	gasoline	False	gasoline	3.6	universal	False	owned	all	17059.12	False	Гродненская о...	24



	body_type	has_warranty	state	drivetrain	price_usd	is_exchangeable	location_region	number_of_photos	up_counter	feature_0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	duration_list...
1	universal	False	owned	all	10900	False	Минская обл.	9	13	False	True	True	True	False	True	False	True	True	True	16
2	universal	False	owned	all	5000	True	Минская обл.	12	54	False	True	False	False	True	True	False	False	False	True	83
3	suv	False	owned	all	2800	True	Минская обл.	4	72	False	True	False	False	False	False	False	False	True	True	151
4	sedan	False	owned	all	9999	True	Минская обл.	9	42	True	False	False	False	False	False	False	False	False	False	86
5	universal	False	owned	all	2134.11	True	Гомельская обл.	14	7	False	True	False	True	True	False	False	False	False	True	7
6	universal	False	owned	all	14700	True	Минская обл.	20	56	False	True	False	False	False	True	False	True	True	True	67
7	universal	False	owned	all	3000	True	Минская обл.	8	147	False	True	False	False	True	True	False	False	False	True	307
8	sedan	False	owned	all	4500	False	Брестская обл.	7	29	False	True	True	False	False	False	False	False	False	True	73
9	universal	False	owned	all	12900	False	Минская обл.	17	33	False	True	True	True	True	True	True	True	True	True	87
10	universal	False	owned	all	4200	True	Минская обл.	8	11	False	True	True	False	False	True	False	False	False	True	43
11	suv	False	owned	all	6900	True	Минская обл.	14	6	False	True	False	True	False	True	False	True	True	True	11
12	suv	False	owned	all	8350	True	Минская обл.	18	61	False	True	True	True	True	True	False	False	True	True	80
13	suv	False	owned	all	4300	False	Минская обл.	13	2	False	True	False	False	False	False	False	False	False	False	2
14	hatchback	False	owned	all	1666	False	Гомельская обл.	8	94	True	False	False	False	False	False	False	False	False	False	230
15	universal	False	owned	all	8600	True	Минская обл.	24	34	False	True	False	True	False	False	False	False	False	True	63
16	universal	False	owned	all	7300	True	Минская обл.	17	22	False	True	True	False	True	True	False	False	False	True	35
17	universal	False	owned	all	7587.97	True	Минская обл.	7	166	False	True	False	True	True	True	False	False	False	True	468
18	sedan	False	owned	all	10950	False	Минская обл.	12	7	False	True	False	False	False	False	False	False	False	True	21
19	suv	False	owned	all	12700	False	Минская обл.	14	14	False	False	True	True	False	True	True	True	True	True	16
20	suv	False	owned	all	16500	False	Минская обл.	6	34	False	True	True	False	True	True	False	True	True	True	59
21	universal	False	owned	all	8700	False	Гомельская обл.	10	11	False	True	True	False	True	True	True	False	True	True	13
22	suv	False	owned	all	7500	False	Минская обл.	18	18	False	True	False	True	True	True	True	False	True	True	188
23	suv	False	owned	all	8650	False	Минская обл.	30	25	False	True	False	False	True	True	False	False	True	True	37
24	universal	False	owned	all	3500	True	Минская обл.	5	42	False	True	False	False	True	False	False	False	True	True	86
25	universal	False	owned	all	3800	True	Могилевская о...	12	44	False	True	False	False	False	False	False	False	True	False	67
26	suv	False	owned	all	7200	True	Гомельская обл.	7	62	False	True	False	False	True	True	False	False	True	True	204
27	universal	False	owned	front	1850	True	Минская обл.	18	23	False	True	False	False	False	True	False	False	True	True	71
28	universal	False	owned	all	3800	False	Брестская обл.	12	10	False	True	False	False	True	True	False	False	False	True	21
29	universal	False	owned	all	3000	False	Минская обл.	6	103	False	False	False	False	False	False	False	False	False	True	581
30	sedan	False	owned	all	6200	False	Витебская обл.	10	21	False	True	True	False	False	False	False	False	False	True	48
31	universal	False	owned	all	3700	False	Минская обл.	6	20	False	True	False	False	True	True	False	False	False	True	60
32	universal	False	owned	all	650	False	Гомельская обл.	9	47	False	False	False	False	False	True	False	False	True	False	85
33	universal	False	owned	all	2000	True	Могилевская о...	4	5	False	False	False	False	False	False	False	False	False	True	91
34	sedan	False	owned	all	800	False	Брестская обл.	6	16	False	False	False	False	False	False	False	False	True	True	20
35	sedan	False	owned	all	14950	True	Минская обл.	26	132	True	False	False	False	False	False	False	False	False	False	174
36	hatchback	False	owned	all	5350	False	Минская обл.	13	11	False	True	True	True	False	True	True	True	False	True	62
37	universal	False	owned	all	17059.12	False	Гродненская о...	24	11	False	True	False	True	True	True	True	True	True	True	15

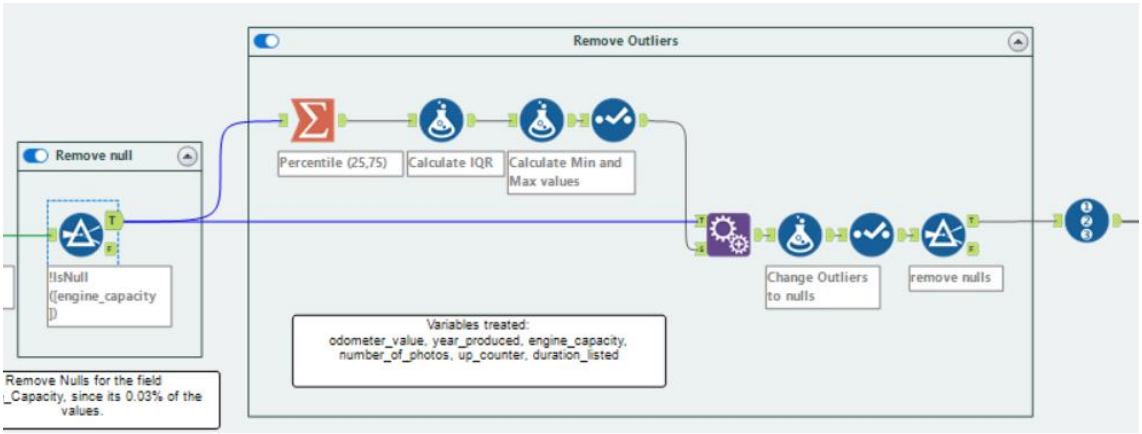
Clean Up:

Nulls:

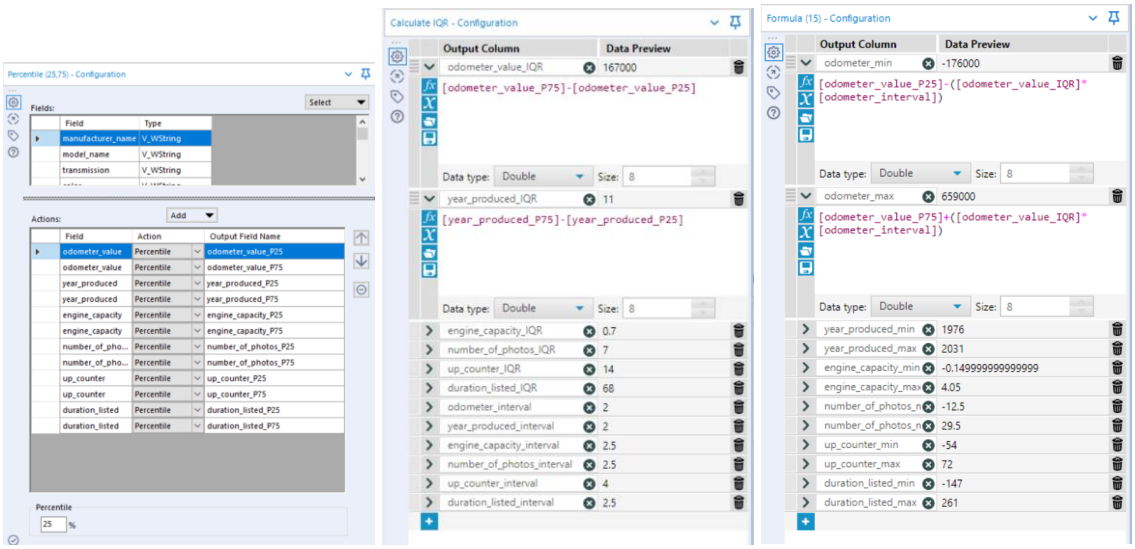


First, we remove the nulls from Engine_Capacity.

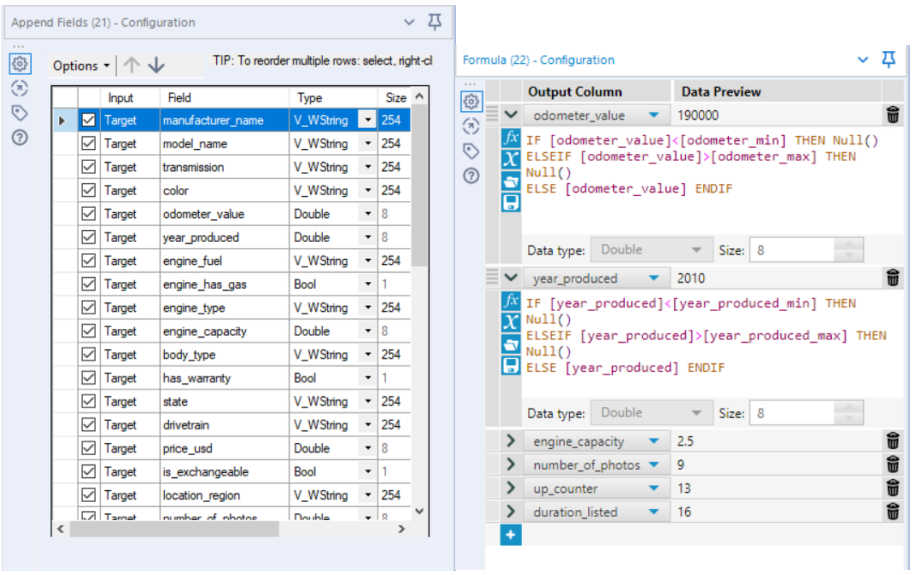
Outliers:



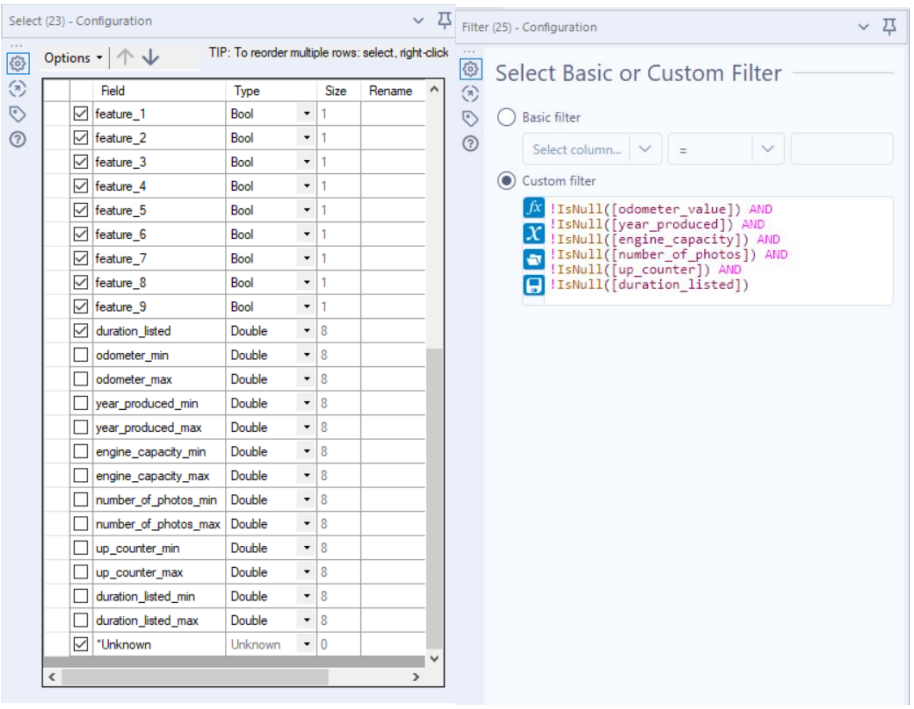
Use Summarize tool to calculate the 25 and 75 percentiles. The Formula Tool to calculate the IQR, and then again the Formula tool to calculate the Upper and Lower limits.



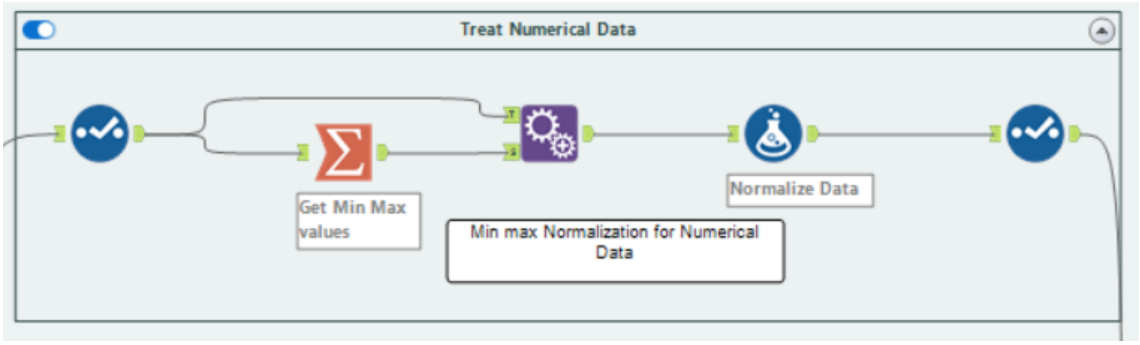
After that we use the Append Fields tool to append the Min and Max limits to the original dataframe. After that we use again the Formula tool to remove the values under the Min and above the Max thresholds.



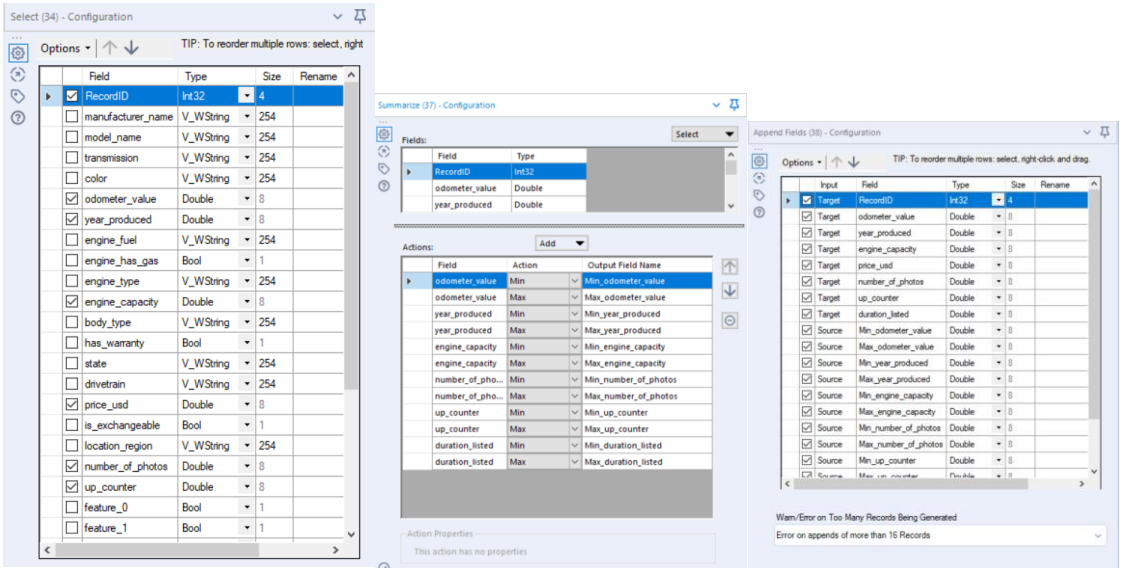
The we use the Select tool, to deselect the Mins and Maxs fields, and a Filter tool to remove the nulls values



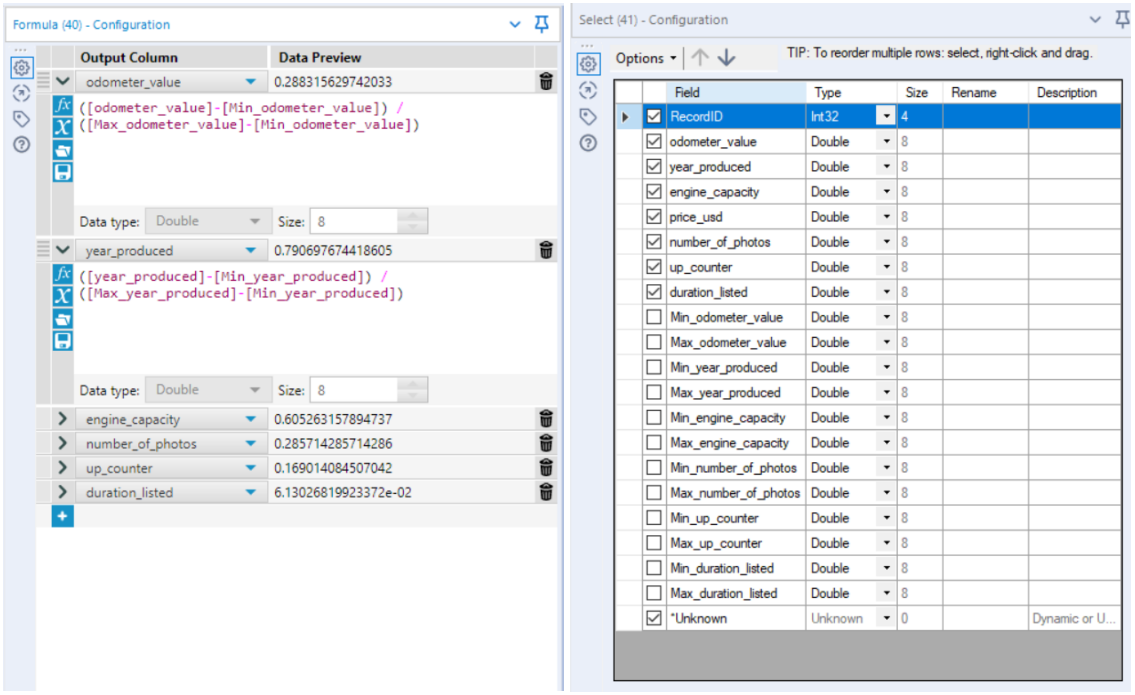
Treat Numerical Data:



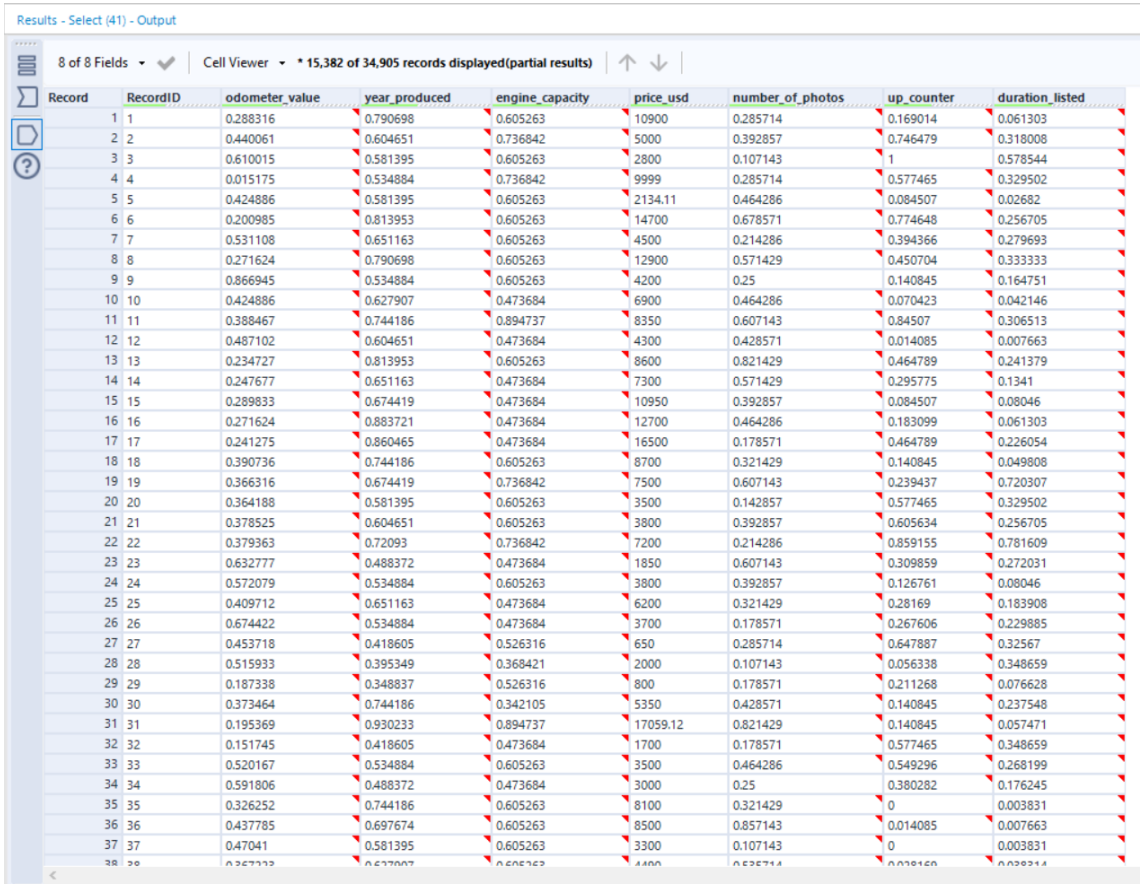
Select tool is used to select only the Numerical Data. Then the Summarize tool is used to get the min and max values of each field. The Append Fields is used to append the min and max values to the dataframe



Lastly the formula tool is used to apply the Min-Max algorithm. Then the Select tool is used to deselect the min, max fields.



We have the next results:

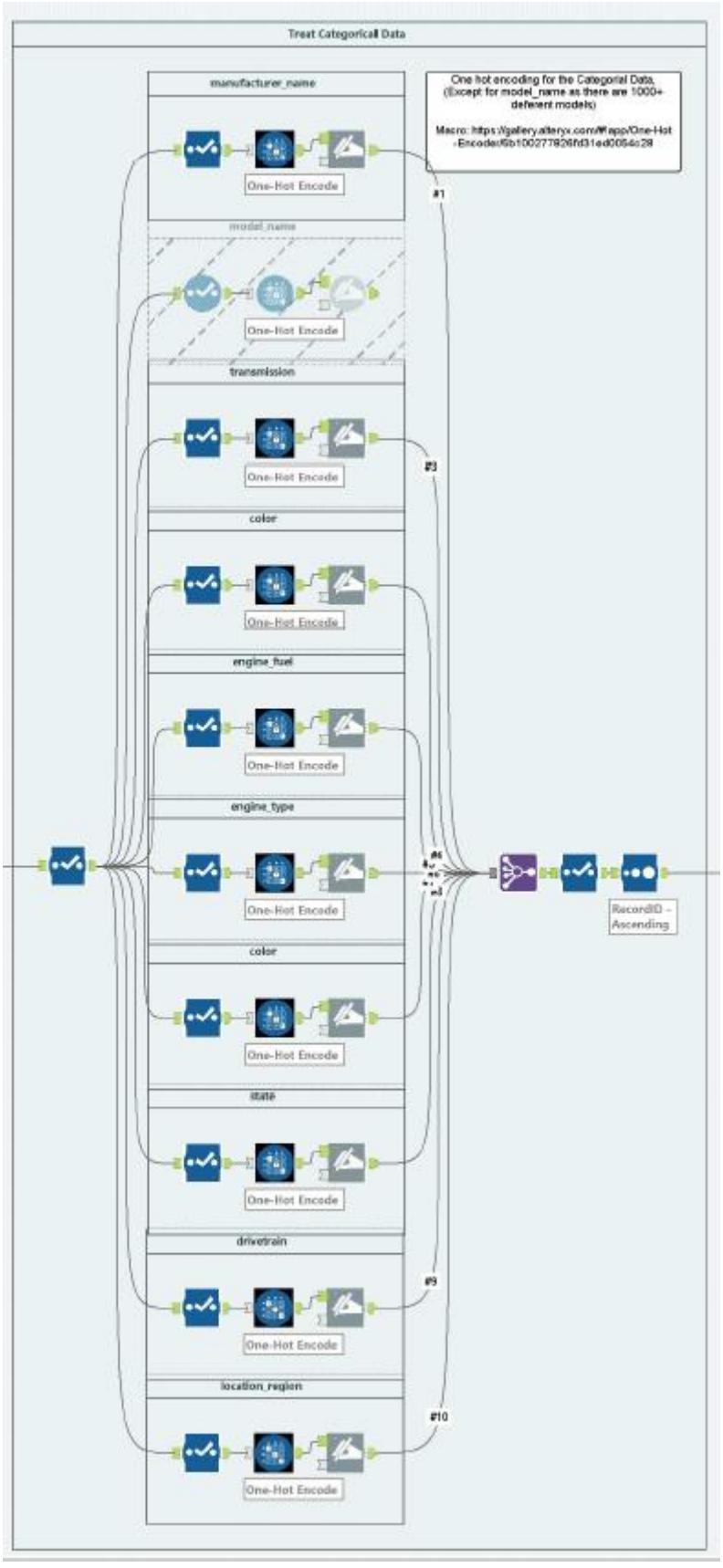


Results - Select (41) - Output

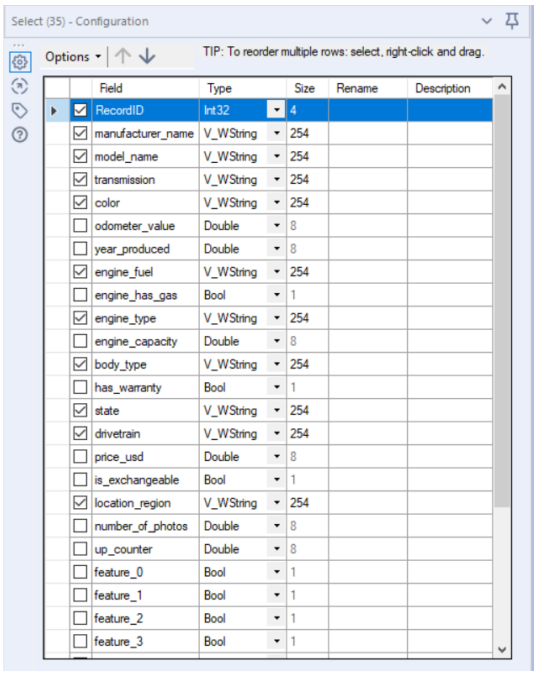
8 of 8 Fields | Cell Viewer | * 15,382 of 34,905 records displayed(partial results)

Record	RecordID	odometer_value	year_produced	engine_capacity	price_usd	number_of_photos	up_counter	duration_listed
1	1	0.288316	0.790698	0.605263	10900	0.285714	0.169014	0.061303
2	2	0.440061	0.604651	0.736842	5000	0.392857	0.746479	0.318008
3	3	0.610015	0.581395	0.605263	2800	0.107143	1	0.578544
4	4	0.015175	0.534884	0.736842	9999	0.285714	0.577465	0.329502
5	5	0.424886	0.581395	0.605263	2134.11	0.464286	0.084507	0.02682
6	6	0.200985	0.813953	0.605263	14700	0.678571	0.774648	0.256705
7	7	0.531108	0.651163	0.605263	4500	0.214286	0.394366	0.279693
8	8	0.271624	0.790698	0.605263	12900	0.571429	0.450704	0.333333
9	9	0.866945	0.534884	0.605263	4200	0.25	0.140845	0.164751
10	10	0.424886	0.627907	0.473684	6900	0.464286	0.070423	0.042146
11	11	0.388467	0.744186	0.894737	8350	0.607143	0.84507	0.306513
12	12	0.487102	0.604651	0.473684	4300	0.428571	0.014085	0.007663
13	13	0.234727	0.813953	0.605263	8600	0.821429	0.464789	0.241379
14	14	0.247677	0.651163	0.473684	7300	0.571429	0.295775	0.1341
15	15	0.289833	0.674419	0.473684	10950	0.392857	0.084507	0.08046
16	16	0.271624	0.883721	0.473684	12700	0.464286	0.183099	0.061303
17	17	0.241275	0.860465	0.473684	16500	0.178571	0.464789	0.226054
18	18	0.390736	0.744186	0.605263	8700	0.321429	0.140845	0.049808
19	19	0.366316	0.674419	0.736842	7500	0.607143	0.239437	0.720307
20	20	0.364188	0.581395	0.605263	3500	0.142857	0.577465	0.329502
21	21	0.378525	0.604651	0.605263	3800	0.392857	0.605634	0.256705
22	22	0.379363	0.72093	0.736842	7200	0.214286	0.859155	0.781609
23	23	0.632777	0.488372	0.473684	1850	0.607143	0.309859	0.272031
24	24	0.572079	0.534884	0.605263	3800	0.392857	0.126761	0.08046
25	25	0.409712	0.651163	0.473684	6200	0.321429	0.28169	0.183908
26	26	0.674422	0.534884	0.473684	3700	0.178571	0.267606	0.229885
27	27	0.453718	0.418605	0.526316	650	0.285714	0.647887	0.32567
28	28	0.515933	0.395349	0.368421	2000	0.107143	0.056338	0.348659
29	29	0.187338	0.348837	0.526316	800	0.178571	0.211268	0.076628
30	30	0.373464	0.744186	0.342105	5350	0.428571	0.140845	0.237548
31	31	0.195369	0.930233	0.894737	17059.12	0.821429	0.140845	0.057471
32	32	0.151745	0.418605	0.473684	1700	0.178571	0.577465	0.348659
33	33	0.520167	0.534884	0.605263	3500	0.464286	0.549296	0.268199
34	34	0.591806	0.488372	0.473684	3000	0.25	0.380282	0.176245
35	35	0.326252	0.744186	0.605263	8100	0.321429	0	0.003831
36	36	0.437785	0.697674	0.605263	8500	0.857143	0.014085	0.007663
37	37	0.47041	0.581395	0.605263	3300	0.107143	0	0.003831

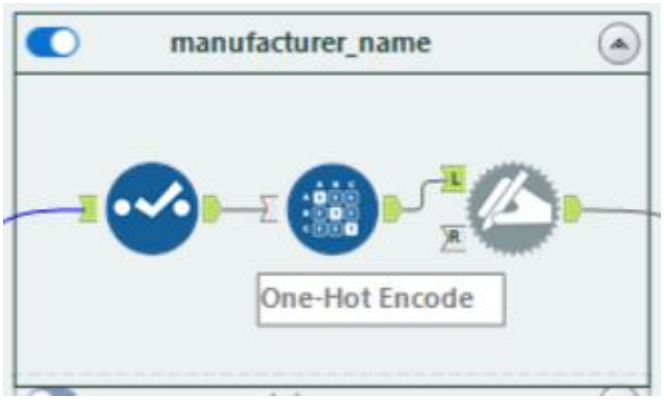
Treat Categorical Data:



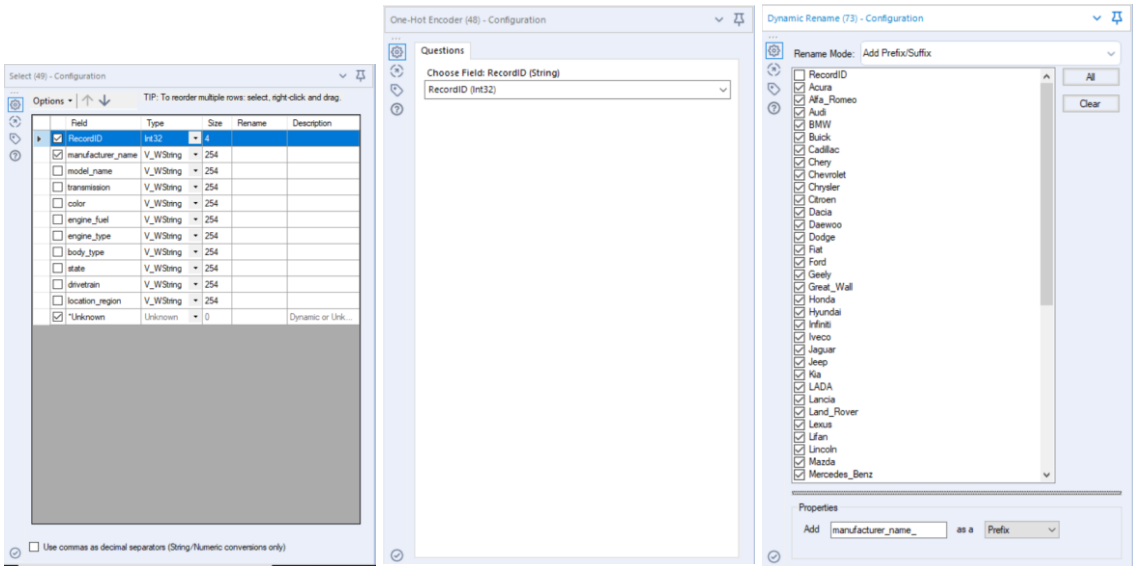
First, we use the Select tool to select only the Categorial variables:



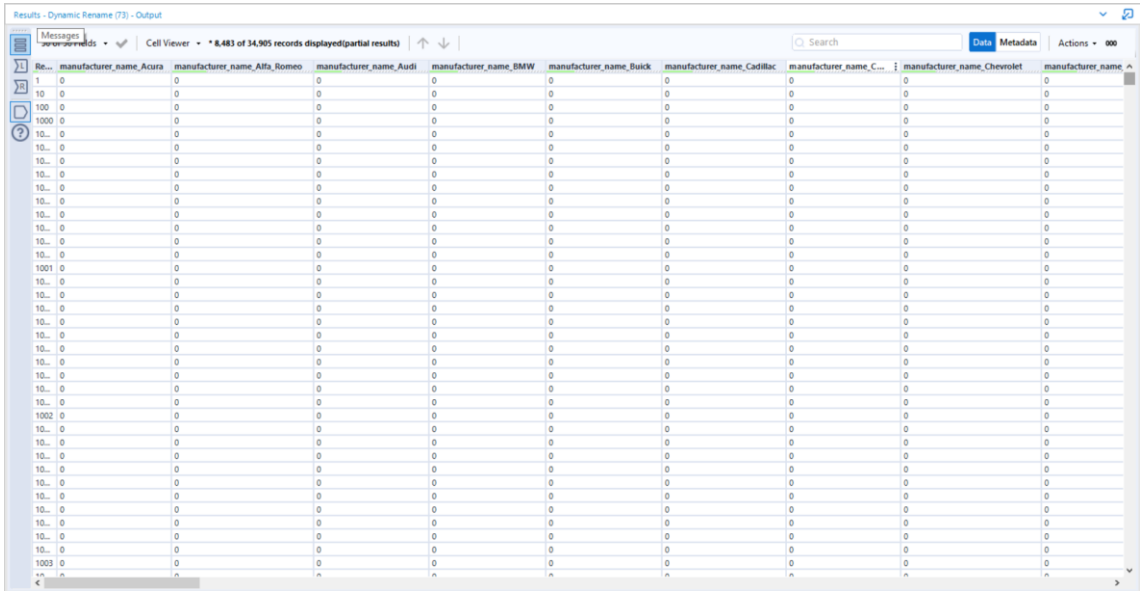
After that each cell do the One-Hot Encoding to each variable, so I'm only going to explain the first one. The variable model_name, is ignore because it has more than a 1000 different values.



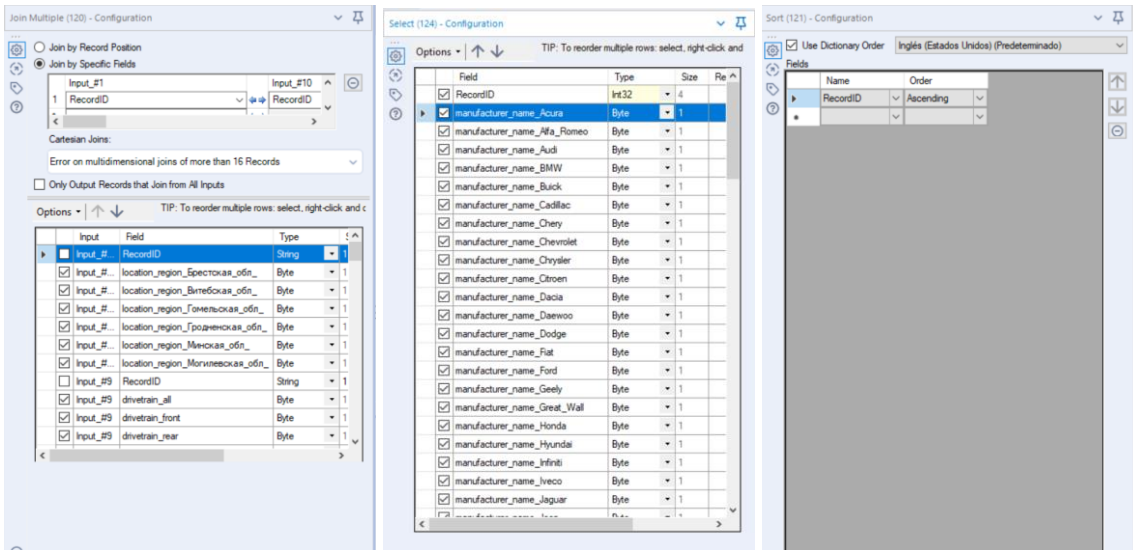
First the select tool is used to select only the RecordID and the manufacturer_name fields. Then we apply the One-Hot Encoding Macro (https://gallery.alteryx.com/#!app/One-Hot-Encoder/5b100277826fd31ed0054c28) And lastly we use the Dynamic Rename tool to add the original name of the field as Prefix



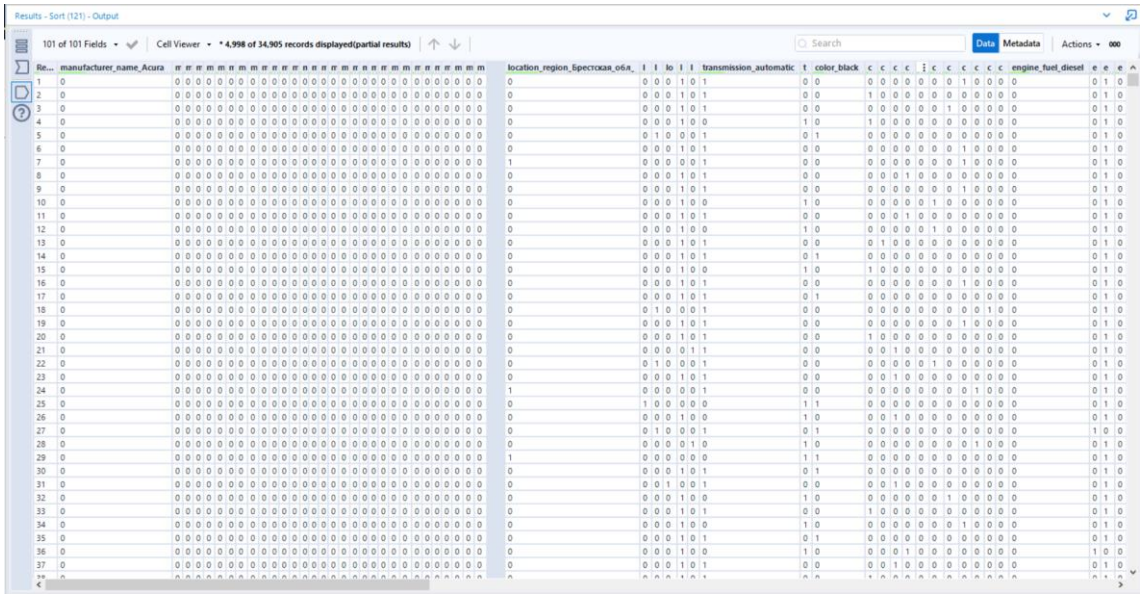
We get the next result:



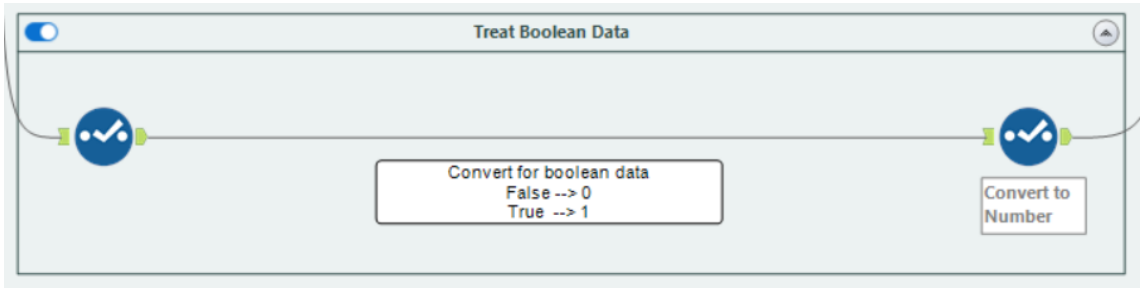
After we complete this process for all the Categorical variables, we Join them using a Join Multiple tool using the RecordID as keys. We then convert the RecordID to an Int, and use a Sort tool to order the Results



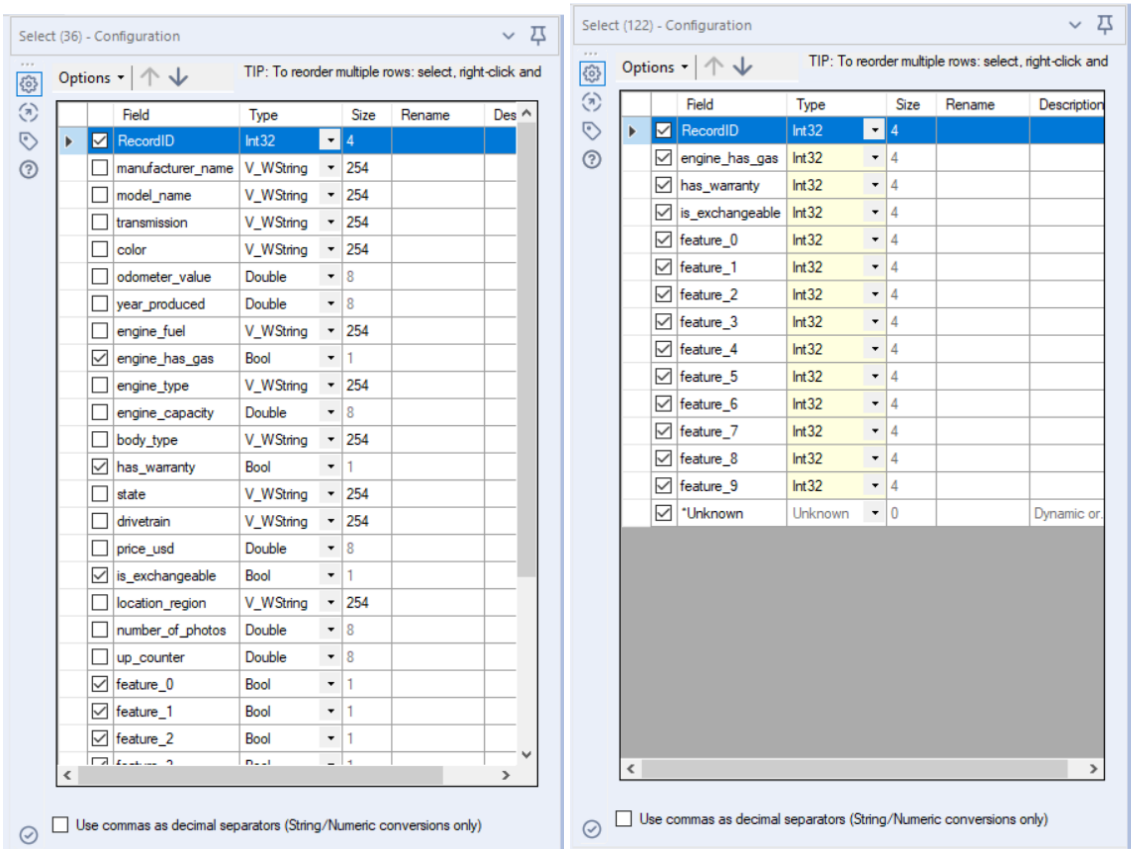
We get the following data:



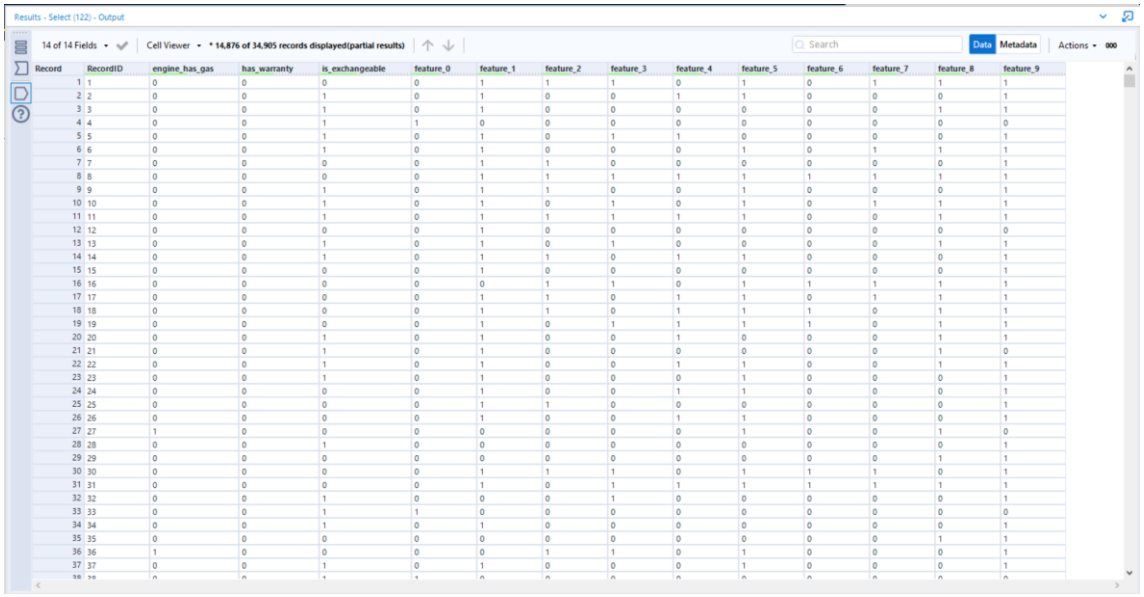
Treat Boolean Data:



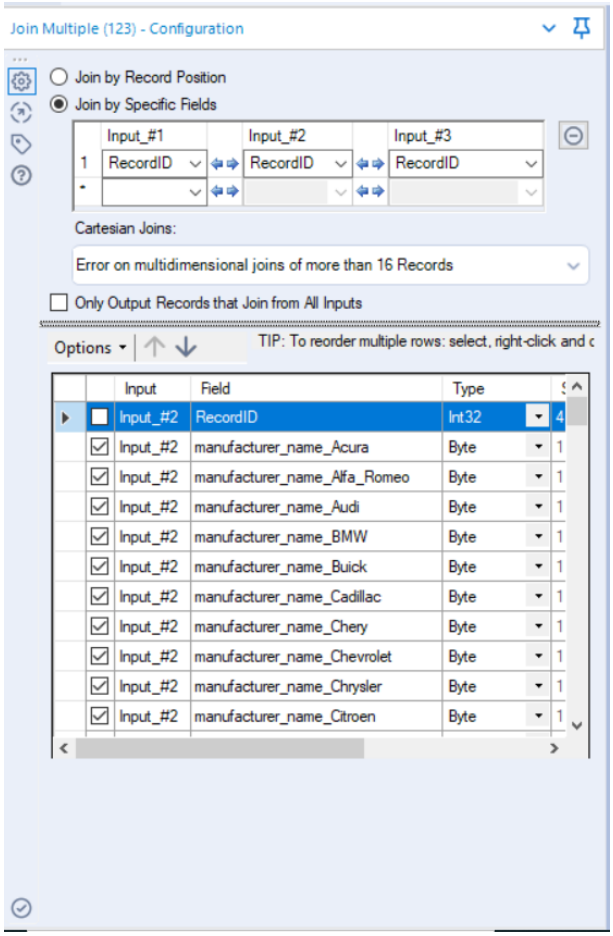
We first select only the Boolean fields using the Select tool. Then we change the types of the variables to Int using the Select tool.



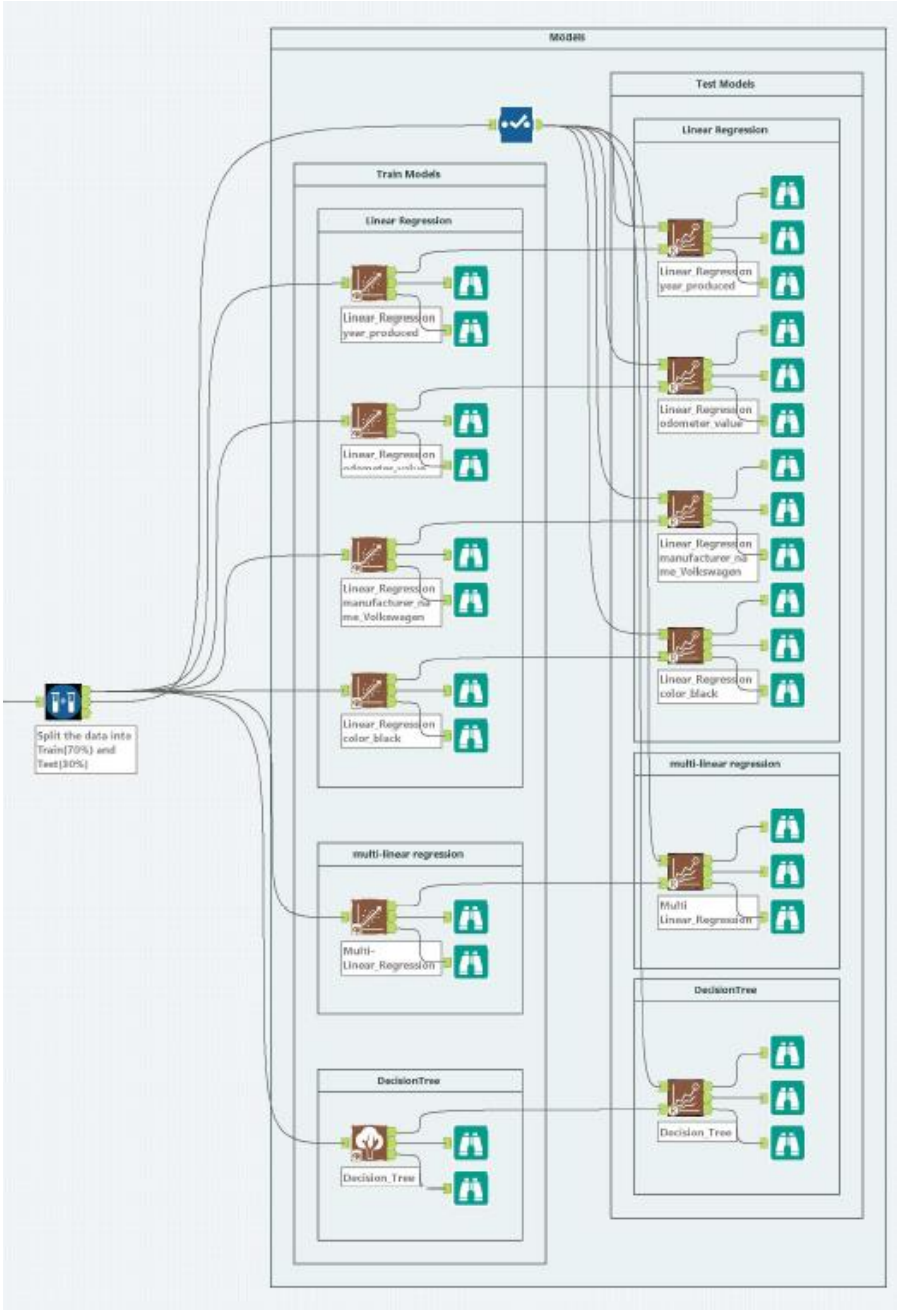
Getting the next results:



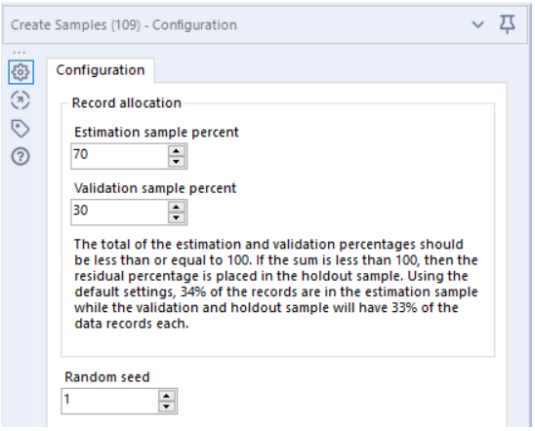
Lastly we Join the outputs from the previous 3 treatment (Numerical, Categorical and Boolean) using a Join Multiple Tool

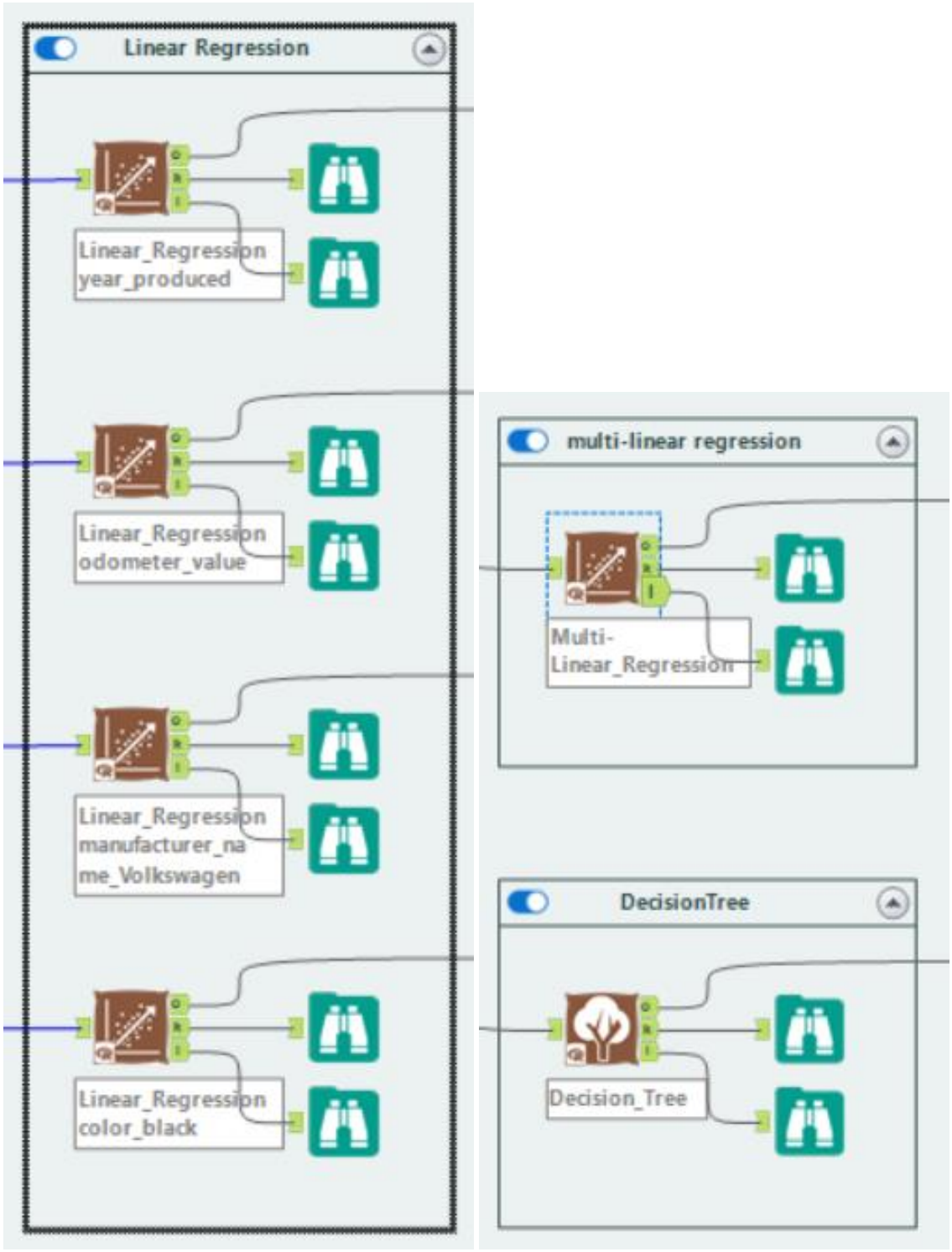


Models:



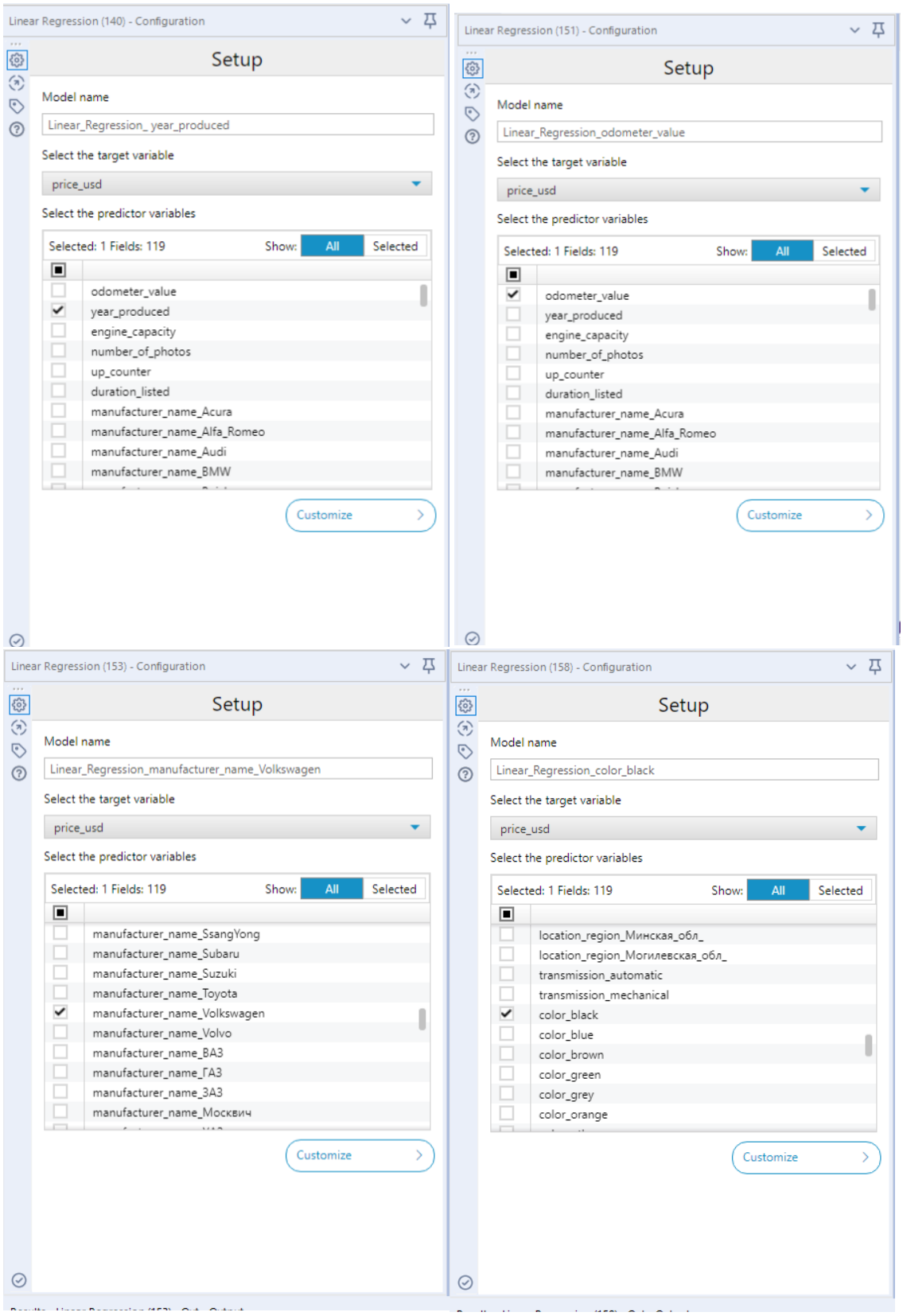
First, we Split the data into Train (70%) and Test (30%) and train the models:



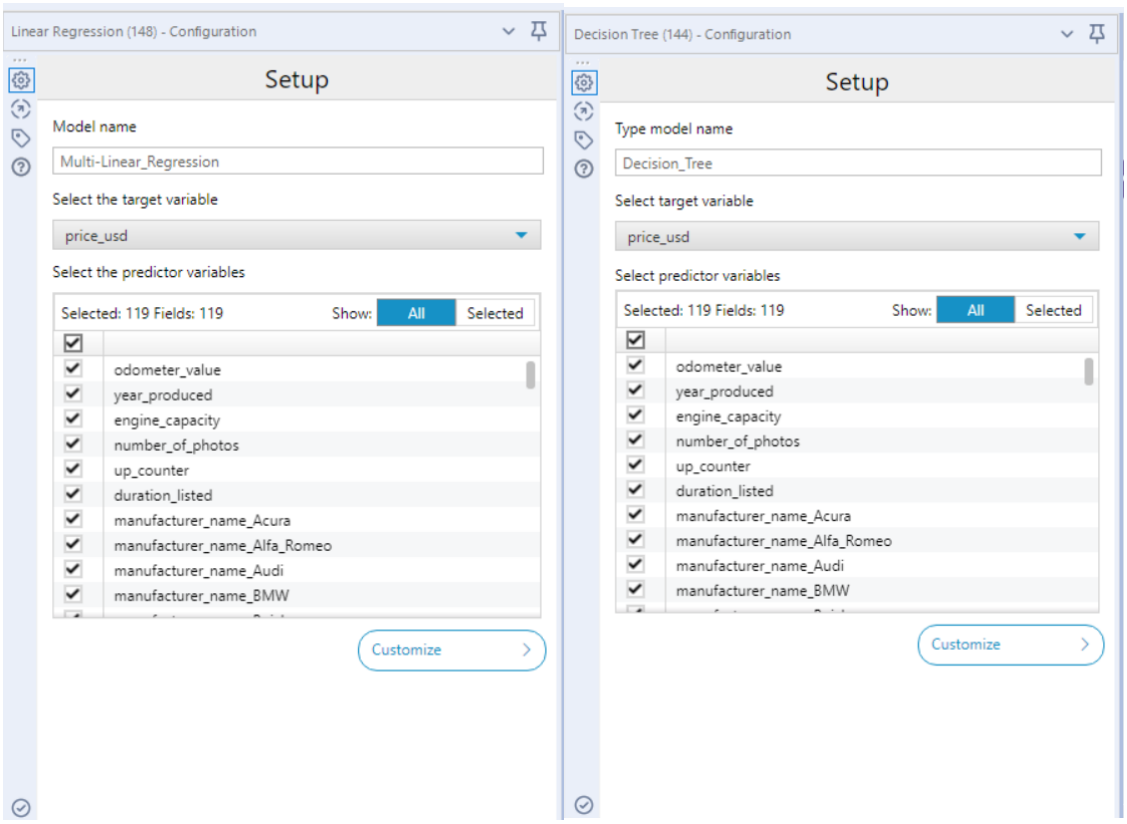

Linear Regression
Linear_Regression
year_produced
Linear_Regression
odometer_value
Linear_Regression
manufacturer_na
me_Volkswagen
Linear_Regression
color_black
multi-linear regression
Multi-
Linear_Regression
DecisionTree
Decision_Tree

Models Train:

Linear regression:



Multi-linear regression & Decision tree:



Results:

Linear regression:

Record Report

Report for Linear Model
Linear_Regression__year_produced

Basic Summary

Call:
lm(formula = price_usd ~ year_produced, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-11501	-2252	-808	1263	37901

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9080	93.76	-96.85	< 2.2e-16 ***
year_produced	24613	143.47	171.55	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4082.1 on 24432 degrees of freedom
Multiple R-squared: 0.5464, Adjusted R-Squared: 0.5464
F-statistic: 29431 on 1 and 24432 degrees of freedom (DF), p-value < 2.2e-16

Record Report

Report for Linear Model
Linear_Regression_odometer_value

Basic Summary

Call:
lm(formula = price_usd ~ odometer_value, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-11959	-3194	-908	2095	40801

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11960	78.75	151.86	< 2.2e-16 ***
odometer_value	-15082	190.80	-79.05	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5408.8 on 24432 degrees of freedom
Multiple R-squared: 0.2037, Adjusted R-Squared: 0.2036
F-statistic: 6248 on 1 and 24432 degrees of freedom (DF), p-value < 2.2e-16

Record	Report
1	**Report for Linear Model** **Linear_Regression_manufacturer_name_Volkswagen**
2	*Basic Summary*
3	Call: lm(formula = price_usd ~ manufacturer_name_Volkswagen, data = the.data)
4	Residuals:

	Min	1Q	Median	3Q	Max
	-6377	-4278	-1778	2209	43622

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6377.59	41.17	154.899	< 2.2e-16	***
manufacturer_name_Volkswagen	-86.56	122.44	-0.707	0.47957	

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6061 on 24432 degrees of freedom
Multiple R-squared: 2.046e-05, Adjusted R-Squared: -2.047e-05
F-statistic: 0.4998 on 1 and 24432 degrees of freedom (DF), p-value 0.4796

Record	Report
1	**Report for Linear Model** **Linear_Regression_color_black**
2	*Basic Summary*
3	Call: lm(formula = price_usd ~ color_black, data = the.data)
4	Residuals:

	Min	1Q	Median	3Q	Max
	-8636	-4007	-1707	2143	44193

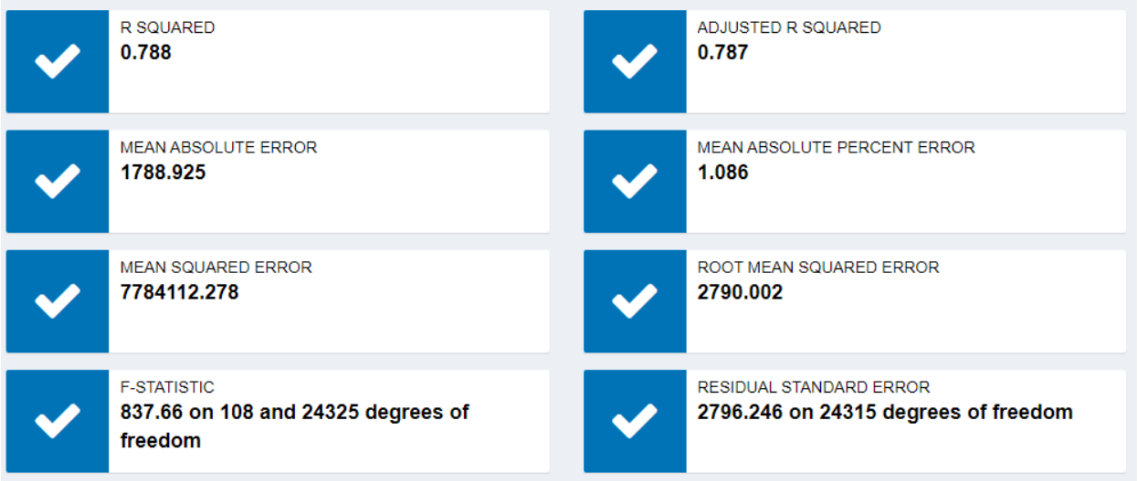
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5807	42.34	137.15	< 2.2e-16	***
color_black	2929	96.73	30.28	< 2.2e-16	***

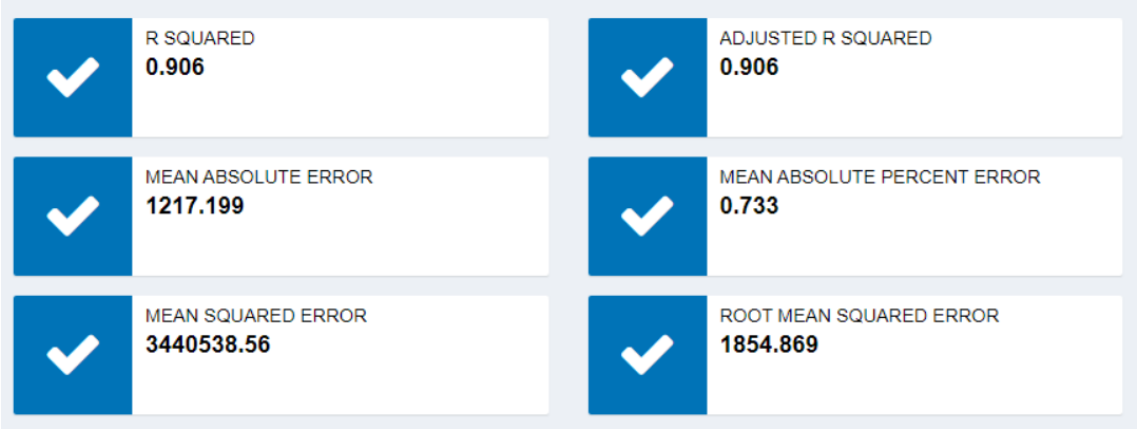
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5950.5 on 24432 degrees of freedom
Multiple R-squared: 0.03617, Adjusted R-Squared: 0.03613
F-statistic: 917 on 1 and 24432 degrees of freedom (DF), p-value < 2.2e-16

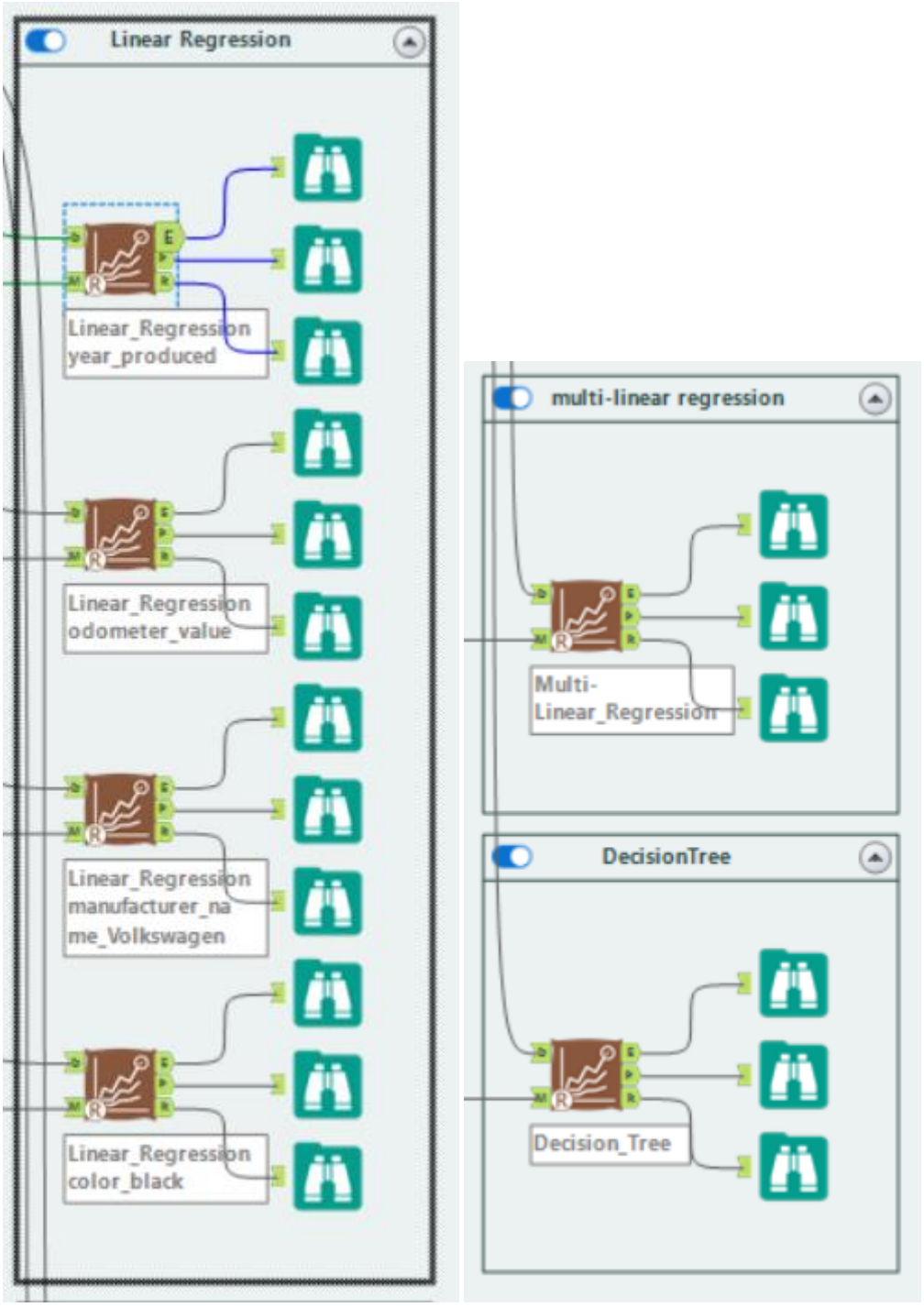
Multi-linear regression:



Decision tree:



Test Models:



Results:

Fit and error measures

Model	Correlation	RMSE	MAE	MPE	MAPE
Linear_Regression_ year_produced	0.7348	4,242.4608	2,729.2587	-21.5418	153.4939

Model	Correlation	RMSE	MAE	MPE	MAPE
Linear_Regression_odometer_value	0.4656	5,535.6134	3,792.1623	-252.2018	285.4971

Model	Correlation	RMSE	MAE	MPE	MAPE
Linear_Regression_manufacturer_name_Volkswagen	0.0031	6,253.4248	4,424.6812	-294.9812	323.8264

Model	Correlation	RMSE	MAE	MPE	MAPE
Linear_Regression_color_black	0.1871	6,143.0067	4,281.5543	-301.5833	329.6564

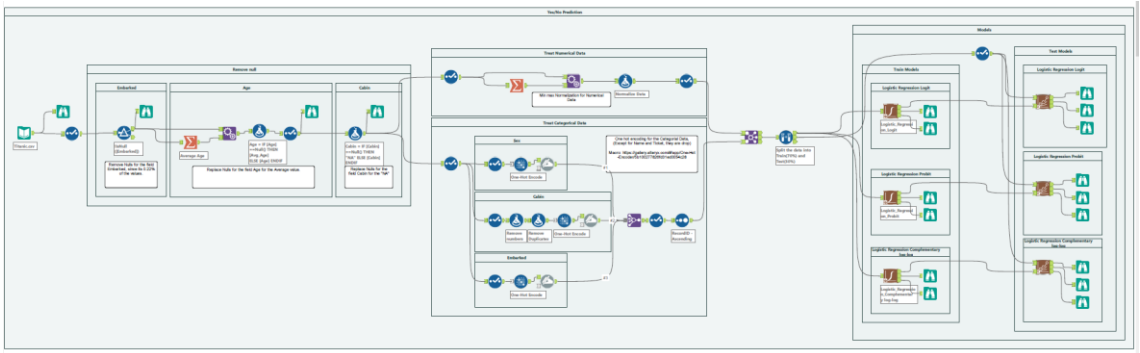
Model	Correlation	RMSE	MAE	MPE	MAPE
Multi-Linear_Regression	0.8812	2,956.0920	1,841.0487	15.5772	73.1644

Model	Correlation	RMSE	MAE	MPE	MAPE
Decision_Tree	0.9336	2,241.3461	1,357.4195	-50.2233	68.0776

Best Model

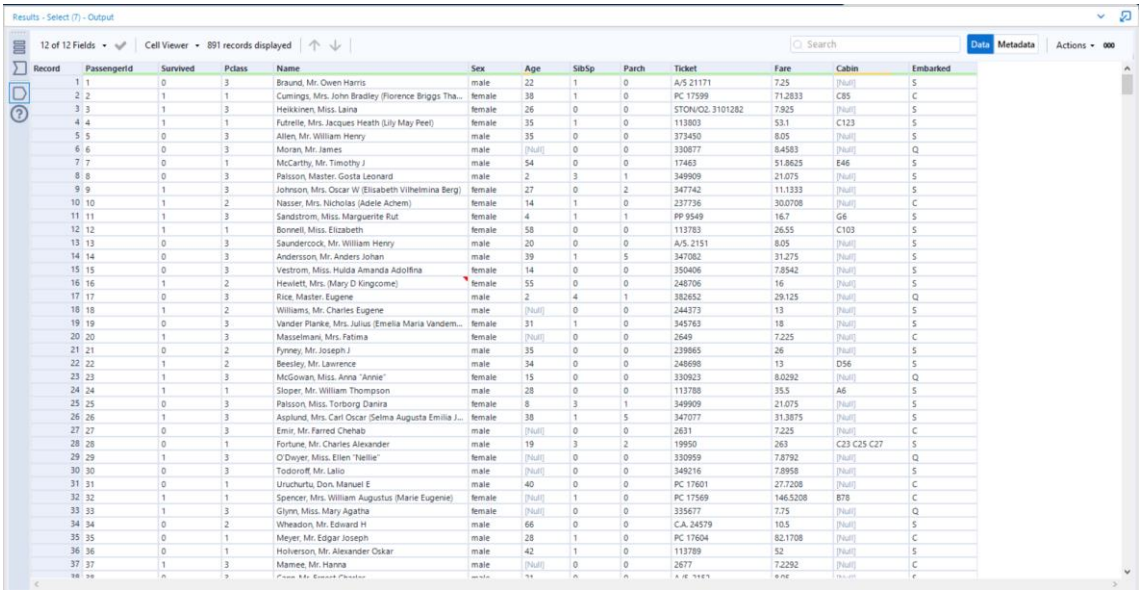
We can see from the result that the model with the best RMSE and MAE is the **Decision Tree**, follow closely by the Multi-Linear Regression. With way worst results we find the different Linear regression Models. We can see the same results if we look at the R2 variable, we find that the Decision Tree have the highest one, follow by the Multi-Linear Regression.

Yes/No Prediction



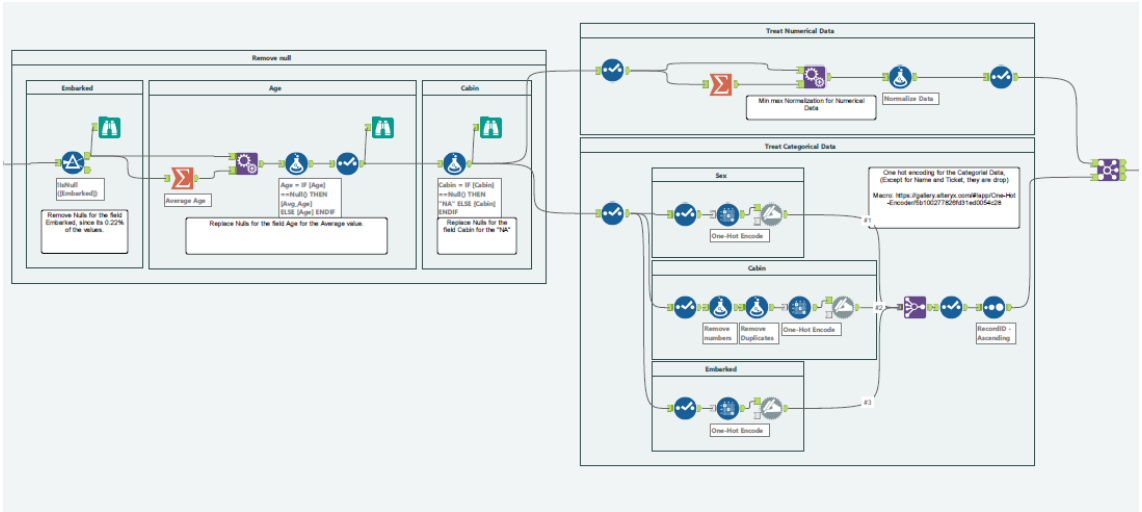
Imported Data:

The imported values are:

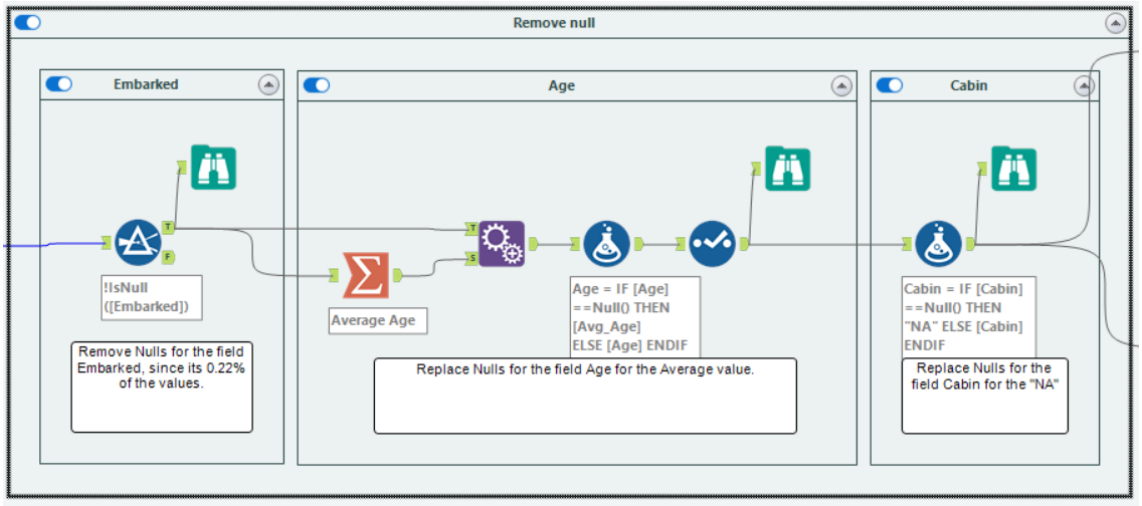


Record	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	[Null]	S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Tha...	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	[Null]	S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	[Null]	S
6	6	0	3	Moran, Mr. James	male	[Null]	0	0	330877	8.4583	[Null]	Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	[Null]	S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	[Null]	S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	[Null]	C
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05	[Null]	S
14	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275	[Null]	S
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542	[Null]	S
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16	[Null]	S
17	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125	[Null]	Q
18	18	1	2	Williams, Mr. Charles Eugene	male	[Null]	0	0	244373	13	[Null]	S
19	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandem...	female	31	1	0	345763	18	[Null]	S
20	20	1	3	Masselmani, Mrs. Fatima	female	[Null]	0	0	2649	7.225	[Null]	C
21	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26	[Null]	S
22	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292	[Null]	Q
24	24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075	[Null]	S
26	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia J...	female	38	1	5	347077	31.3875	[Null]	S
27	27	0	3	Emir, Mr. Farred Chehab	male	[Null]	0	0	2631	7.225	[Null]	C
28	28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
29	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	[Null]	0	0	330959	7.8792	[Null]	Q
30	30	0	3	Todoroff, Mr. Lalio	male	[Null]	0	0	349216	7.8958	[Null]	S
31	31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208	[Null]	C
32	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	[Null]	1	0	PC 17569	146.5208	B78	C
33	33	1	3	Glynn, Miss. Mary Agatha	female	[Null]	0	0	335677	7.75	[Null]	Q
34	34	0	2	Wheadon, Mr. Edward H	male	66	0	0	C.A. 24579	10.5	[Null]	S
35	35	0	1	Meyer, Mr. Edgar Joseph	male	28	1	0	PC 17604	82.1708	[Null]	C
36	36	0	1	Holverson, Mr. Alexander Oskar	male	42	1	0	113789	52	[Null]	S
37	37	1	3	Mamee, Mr. Hanna	male	[Null]	0	0	2677	7.2292	[Null]	C

Clean Up:

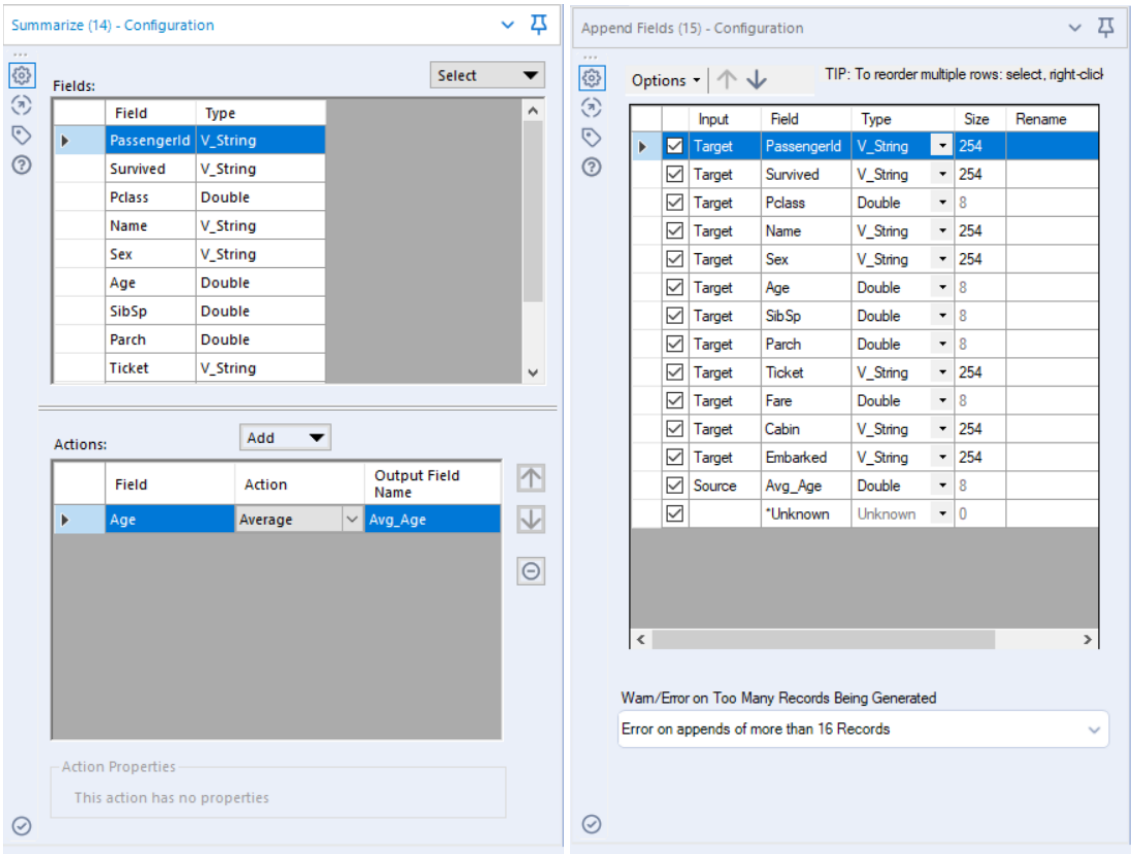


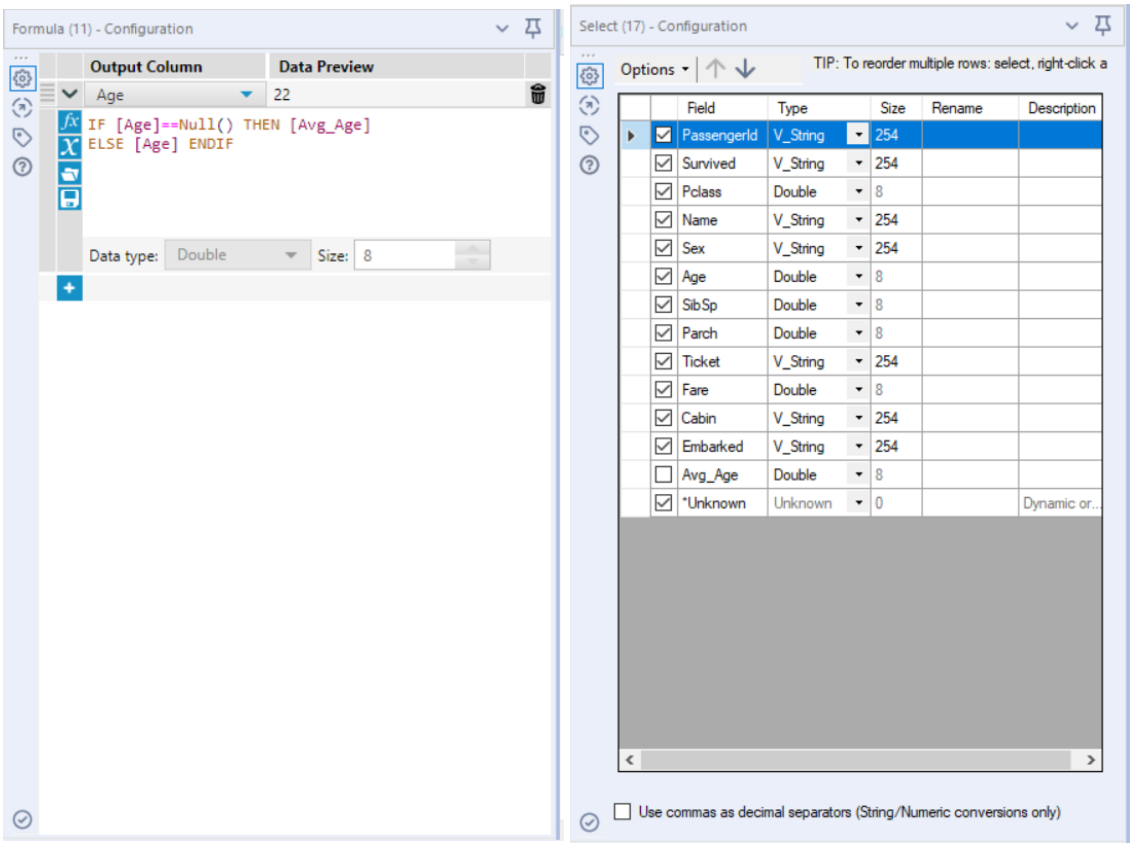
Nulls:



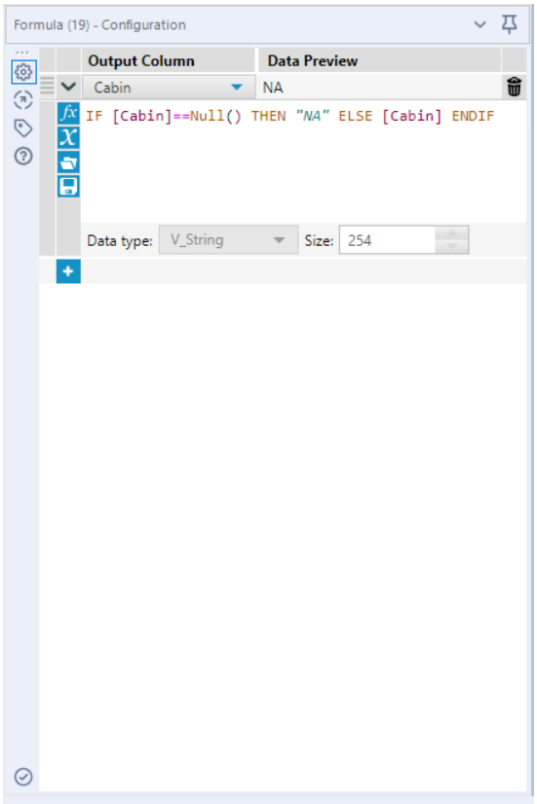
First, we remove the Nulls from Embarked using a Filter tool.

For Age we calculate the Average Using the Summarize tool, then using the Append Field we add it to the original Dataframe, and with the formula tool we change the nulls for the Average. Then the Select tool is used to deselect the Average field

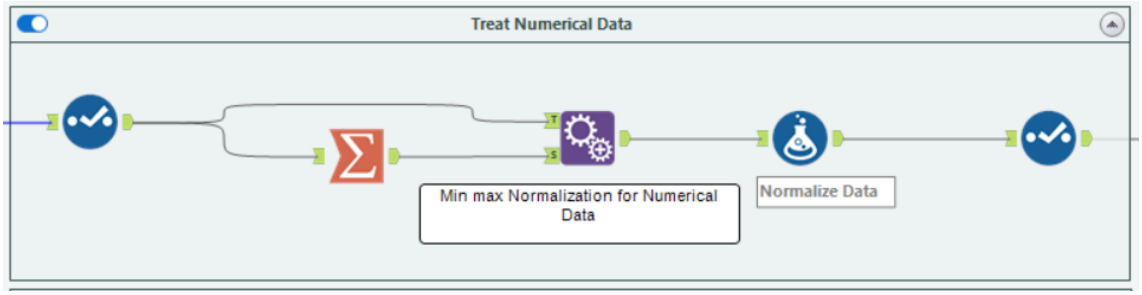




For Cabin, we replace the nulls with the constant NA

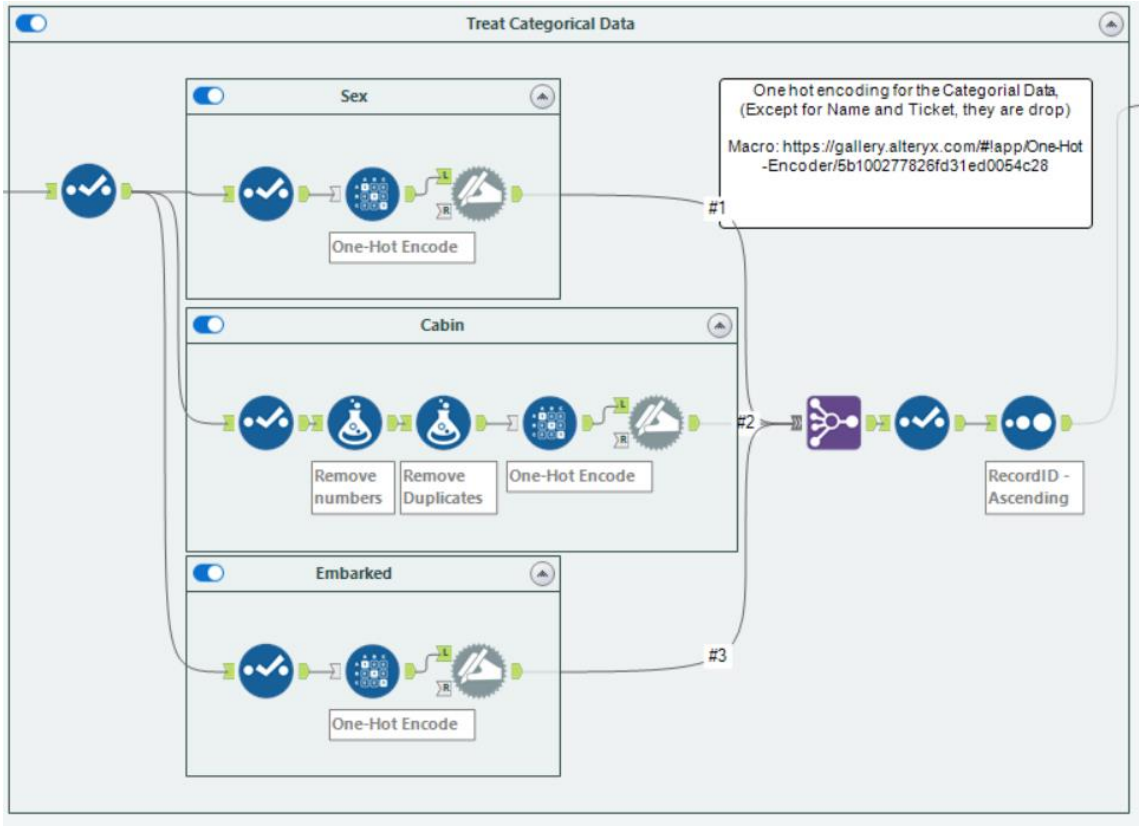


Treat Numerical Data:



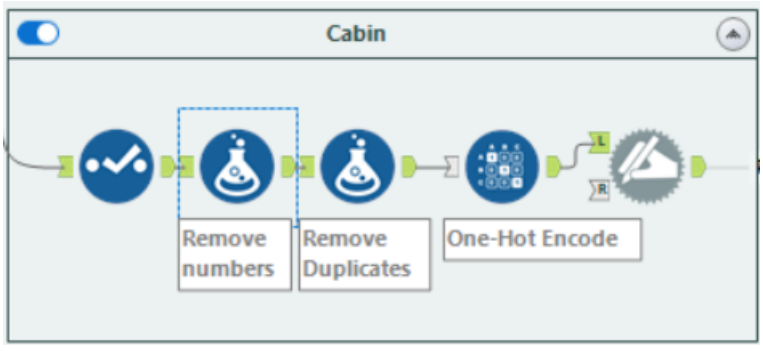
For Normalizing the Numerical Data, we used the same process used in the Numerical Predictor.

Treat Categorical Data:

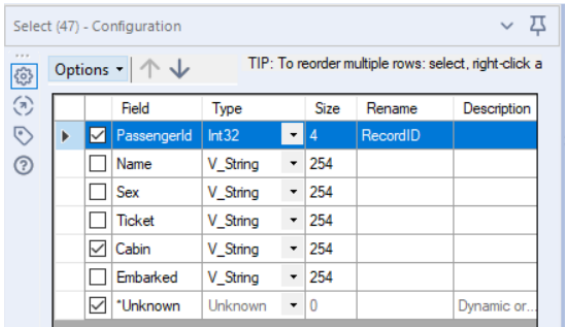


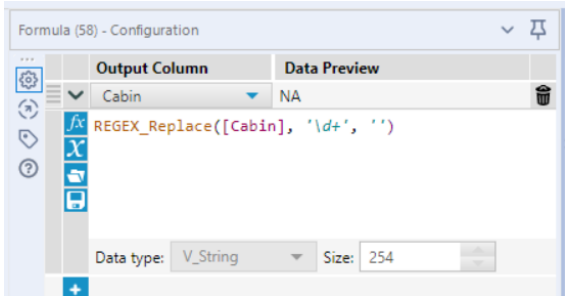
For the Categorical Data, the variables Name and Ticket are drop because they don't bring any useful information. For the Sex and Embarked the same process from the Numerical Predictor is used.

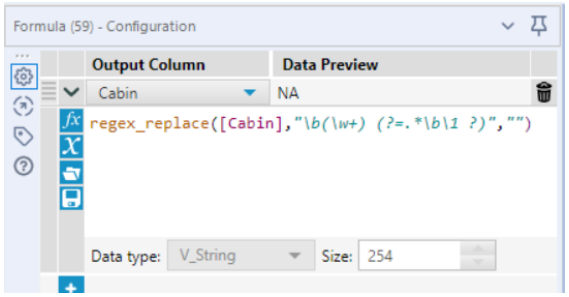
The only difference is in Cabin, where we prepare the data before using the One-Hot encoding.



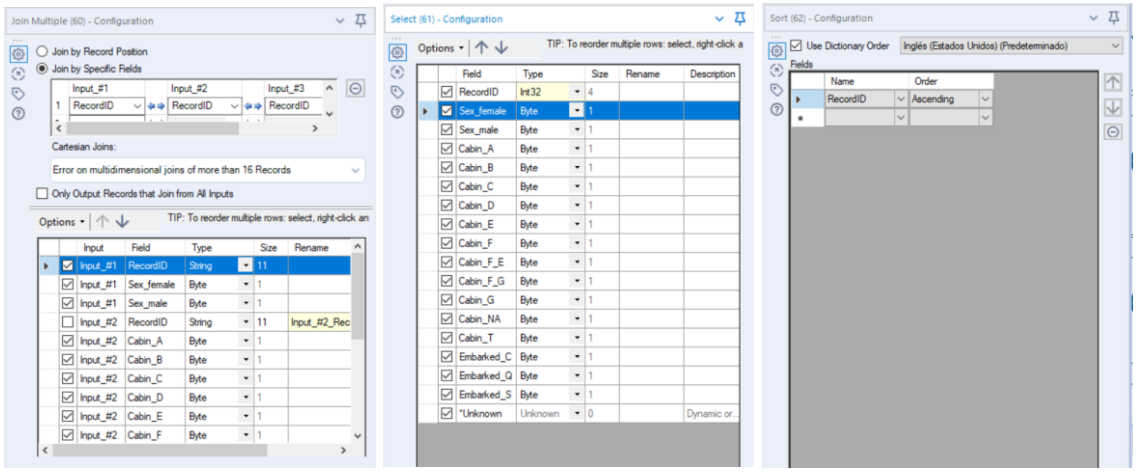
Fist we use the select tool to select only the PassangerID and the Cabin fields. Then the Formula tool is use remove the numbers for the cabins, and a second Formula tool is used to delete duplicate letters, then we perform the One-Hot Encoding like before.



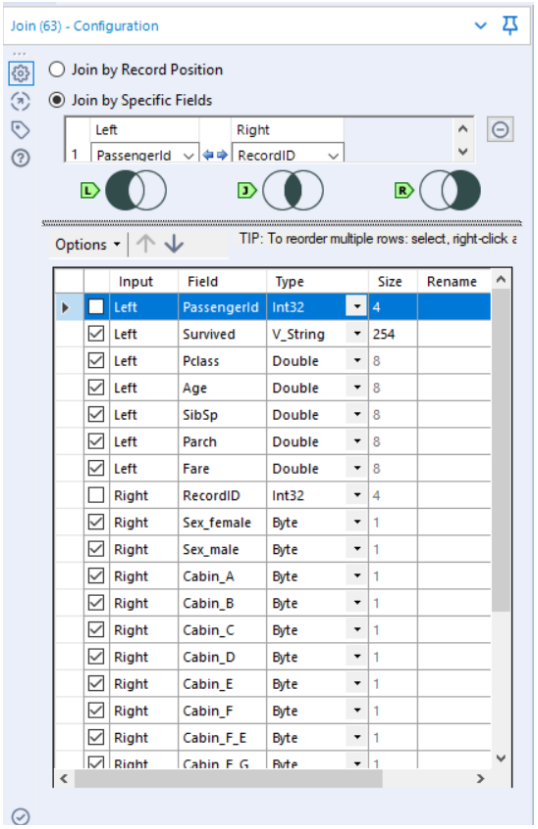




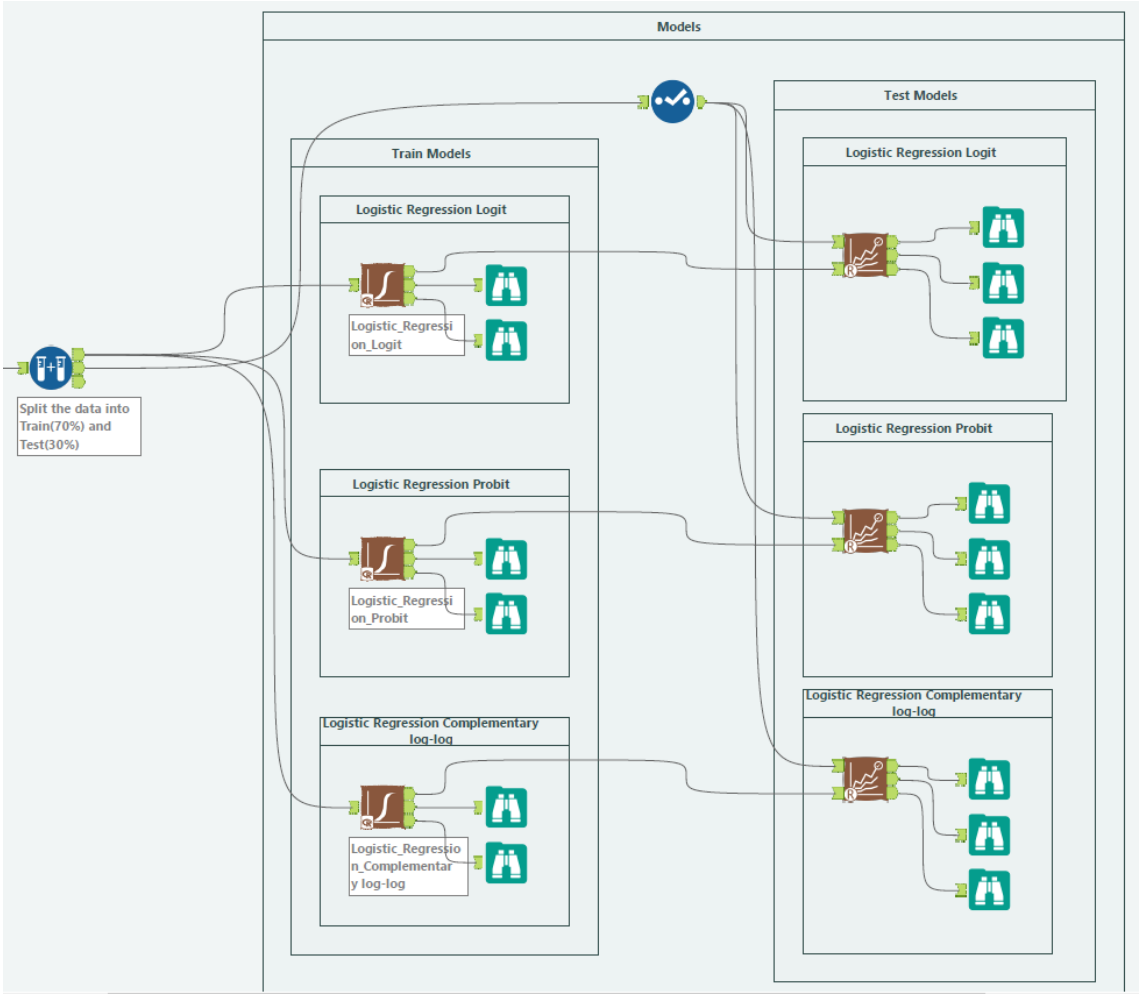
As on the Numerical Predictor, we Join all the categorical variables using the Multiple Join, then the Select tool change the type of RecordID to int, lastly the Sort tool is used to order.



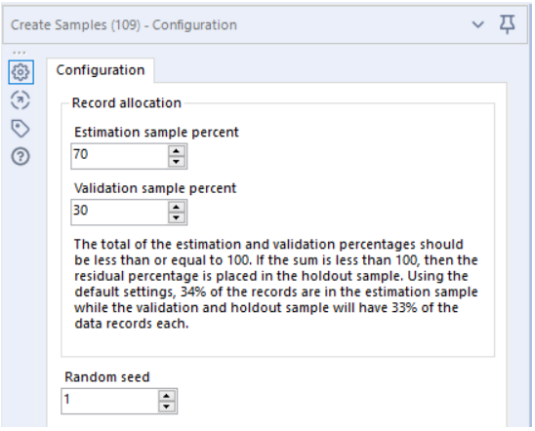
Finaly a Join tool is used to join the outputs from the Numerical and Categorical treatments



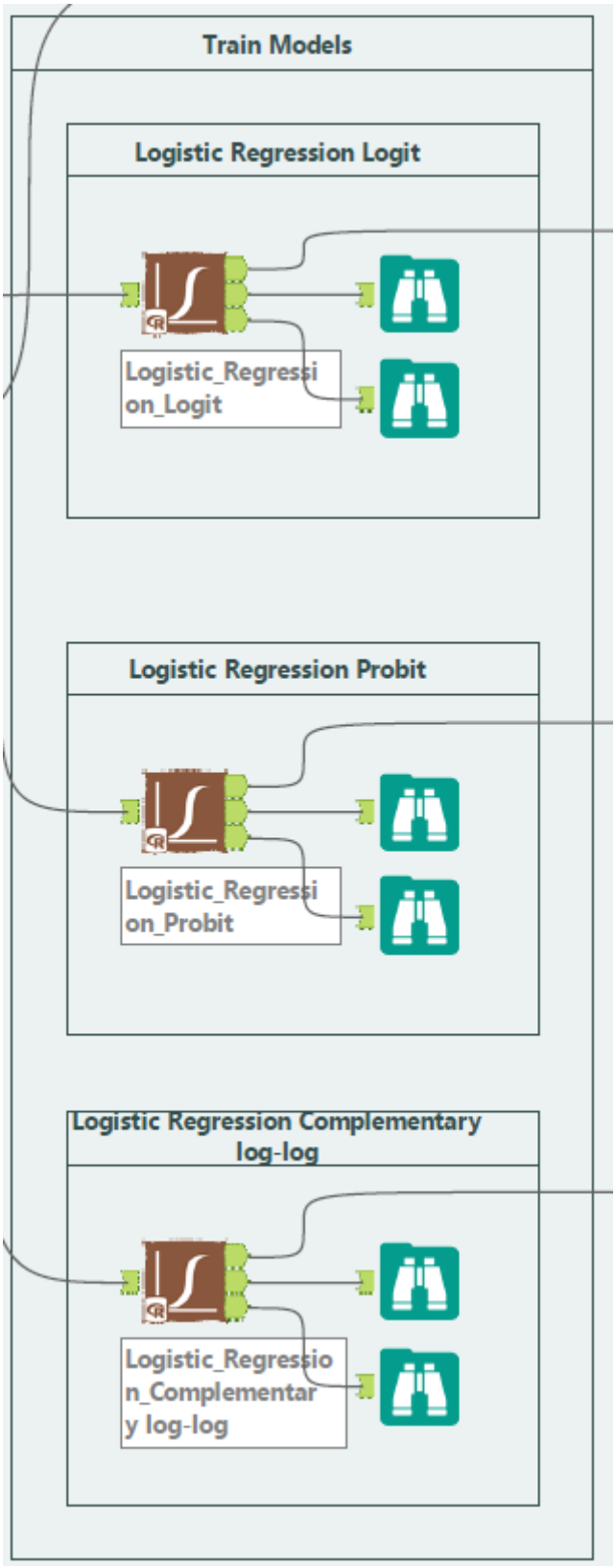
Models:



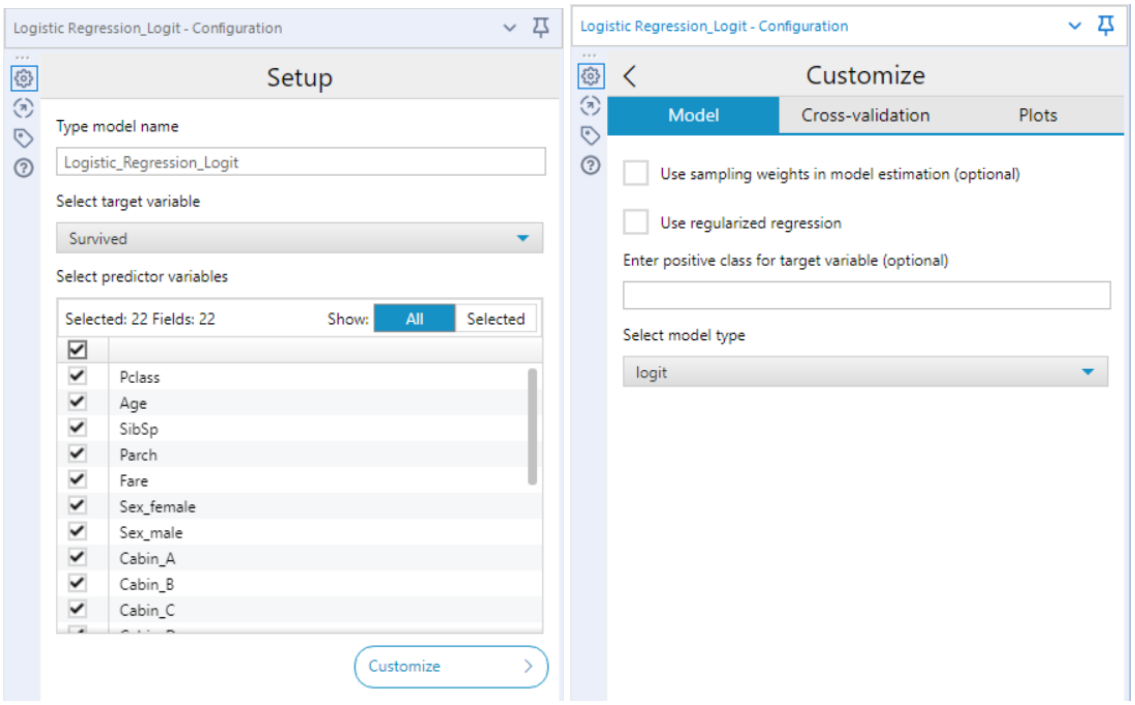
First, we Split the data into Train (70%) and Test (30%) and train the models:



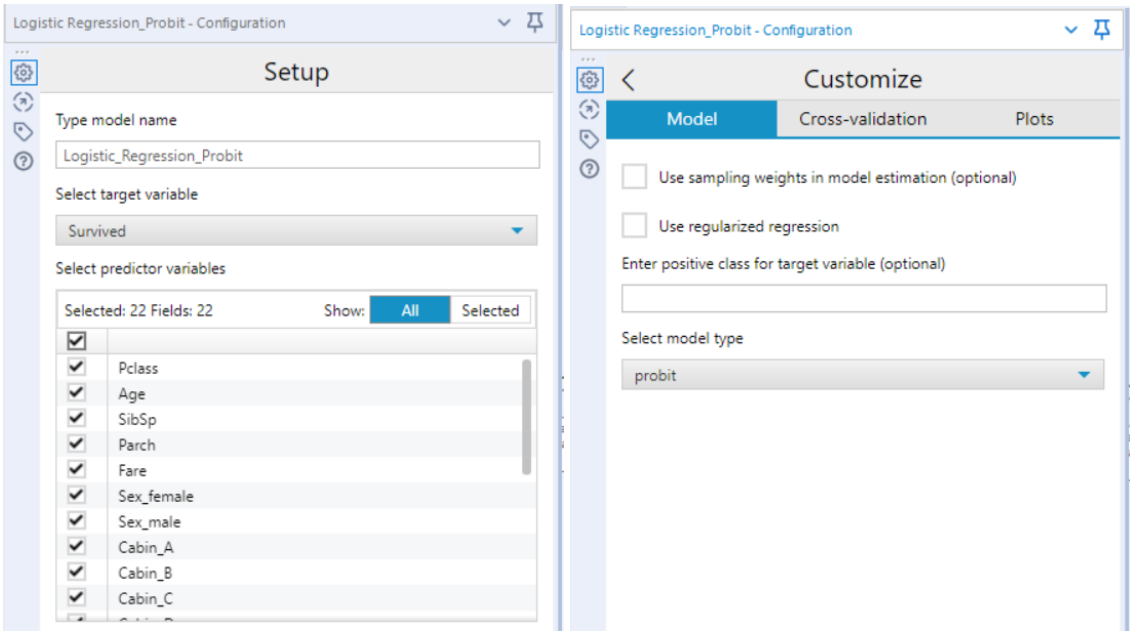
Models Trainc:



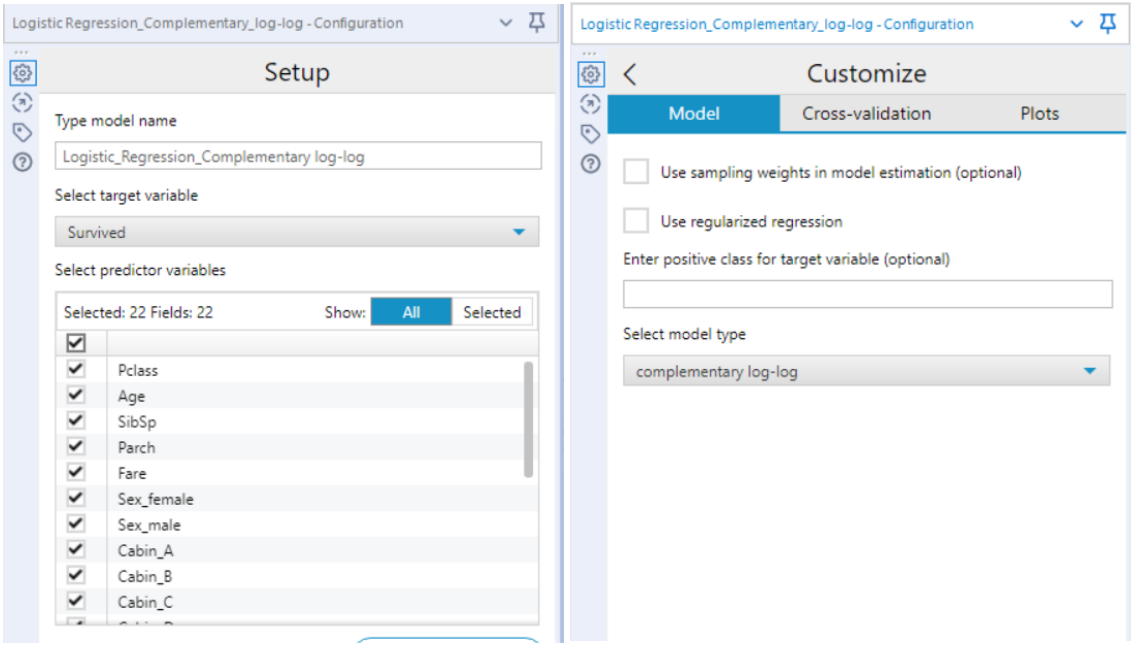
Logistic Regression Logit:



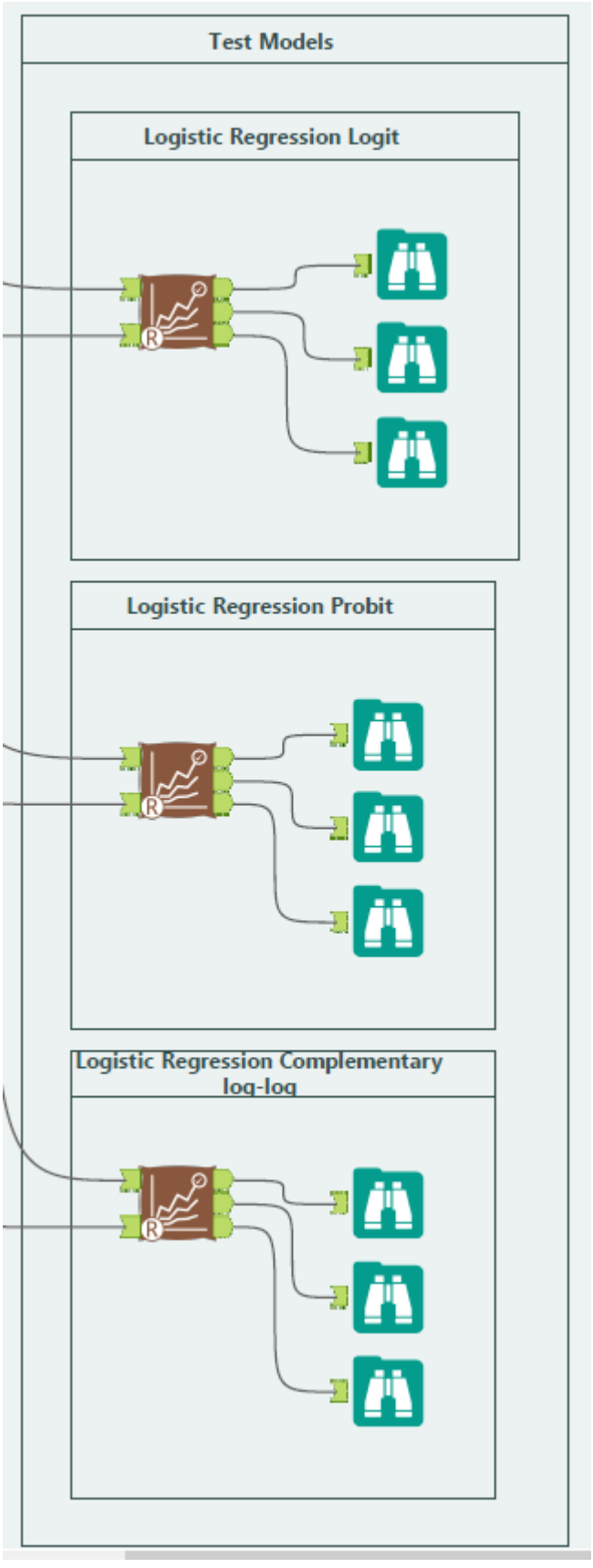
Logistic Regression Probit:



Logistic Regression Complementary log-log:



Test Models:


Test Models
Logistic Regression Logit
Logistic Regression Probit
Logistic Regression Complementary log-log

Results:

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logistic_Regression_Logit	0.8202	0.8596	0.8539	0.9074	0.6857

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logistic_Regression_Probit	0.8165	0.8563	0.8543	0.9012	0.6857

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logistic_Regression_Complementary log-log	0.8315	0.8703	0.8592	0.9321	0.6762

Confusion matrix of Logistic_Regression_Logit

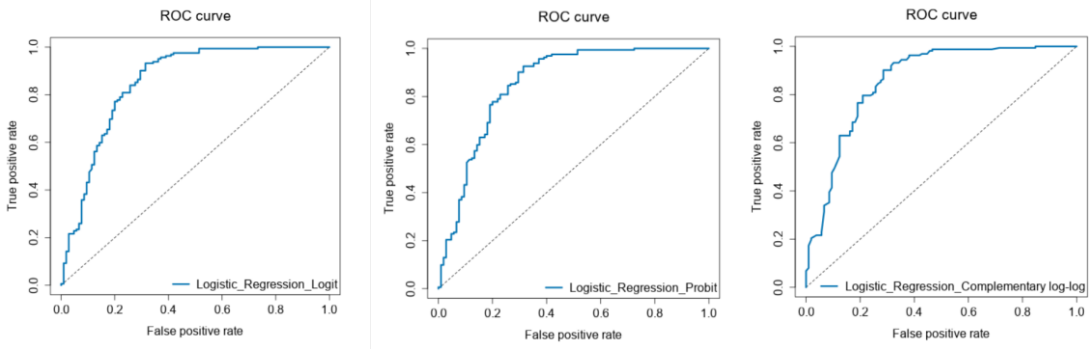
	Actual_0	Actual_1
Predicted_0	147	33
Predicted_1	15	72

Confusion matrix of Logistic_Regression_Probit

	Actual_0	Actual_1
Predicted_0	146	33
Predicted_1	16	72

Confusion matrix of Logistic_Regression_Complementary log-log

	Actual_0	Actual_1
Predicted_0	151	34
Predicted_1	11	71



Best Model

As we can see the **Complementary Log-Log** model perform the best with an AUC value of 0.8592, follow closely by the Probit model with AUC of 0.8543, and lastly the Logit model with an AUC of 0.08539. We can also see that the Complementary Log-Log is the best model from the Confusion Matrix.