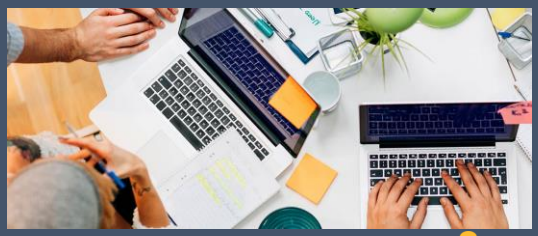


Data Science Project Report

Criselyn Santiago
MoTIS-9

Walmart: Store Sales Forecasting

The data which formed part of my project was the Walmart dataset obtained from Kaggle. The data contained weekly sales of various departments within different stores over period of time. Most of the work put into the project

I mentioned three steps in my proposal, same as the other projects, then, I can proceed also to those steps. Those steps are data cleaning, predictive modeling and visualization. Let's start.

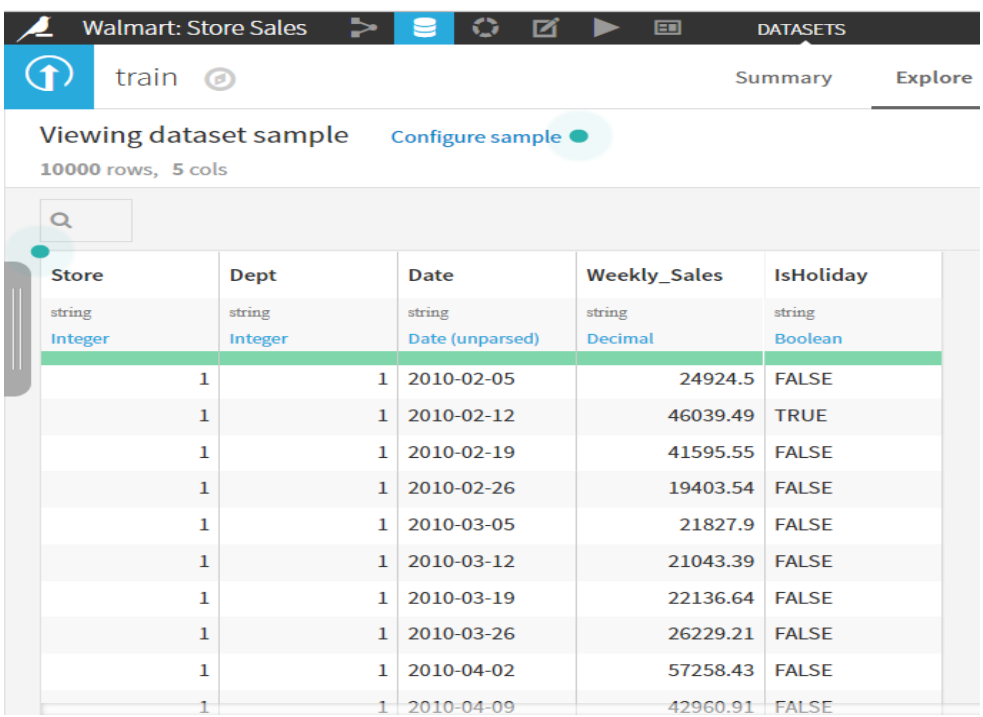
- Data Cleaning

Cleaning process of our data and building some features will be done in data cleaning. In this part, I will do some transformation in my dataset using some recipes like prepare, group, join, enrich, sort, etc. in Dataiku DSS. In this step, I need to analyze my dataset in order to clean it properly and to apply some features which is useful for the dataset. In my case, I just used prepare and join recipe.

- Here are my datasets:

TRAIN DATASET

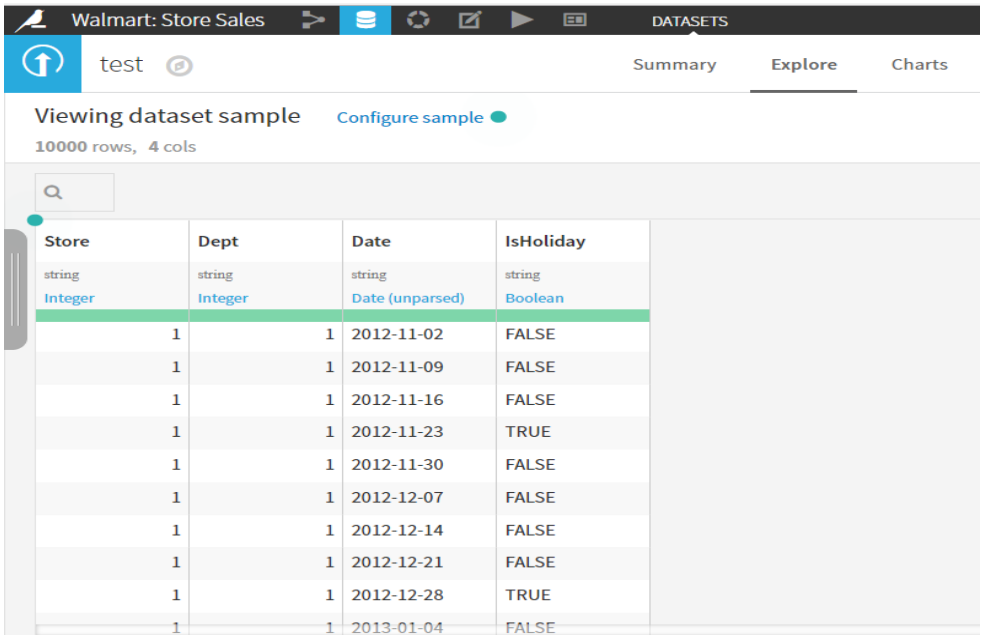
My train dataset is clean as what we noticed in the picture below. So, I don't need to clean it.



Store	Dept	Date	Weekly_Sales	IsHoliday
string Integer	string Integer	string Date (unparsed)	string Decimal	string Boolean
1	1	2010-02-05	24924.5	FALSE
1	1	2010-02-12	46039.49	TRUE
1	1	2010-02-19	41595.55	FALSE
1	1	2010-02-26	19403.54	FALSE
1	1	2010-03-05	21827.9	FALSE
1	1	2010-03-12	21043.39	FALSE
1	1	2010-03-19	22136.64	FALSE
1	1	2010-03-26	26229.21	FALSE
1	1	2010-04-02	57258.43	FALSE
1	1	2010-04-09	42960.91	FALSE

TEST DATASET

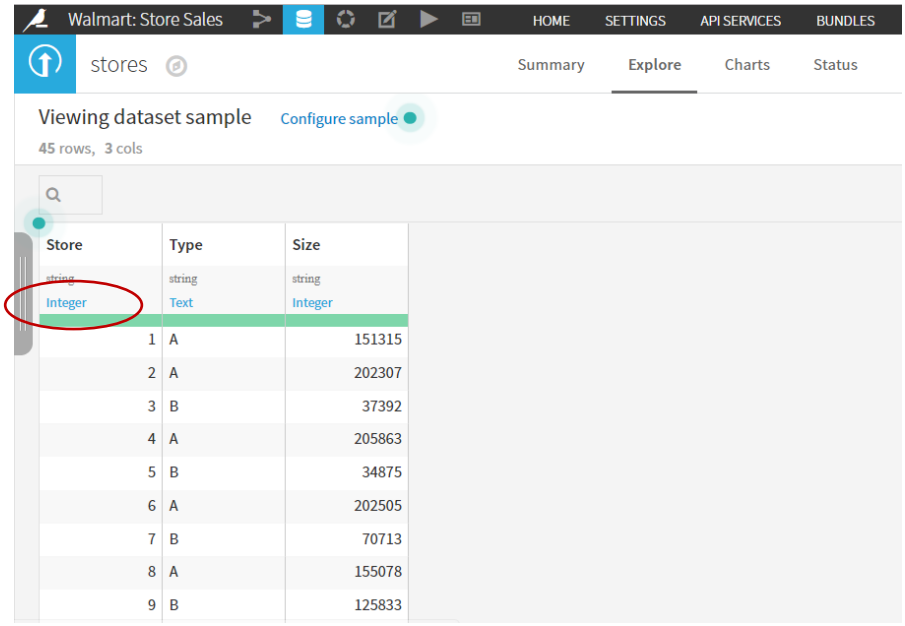
Same thing with the train dataset.



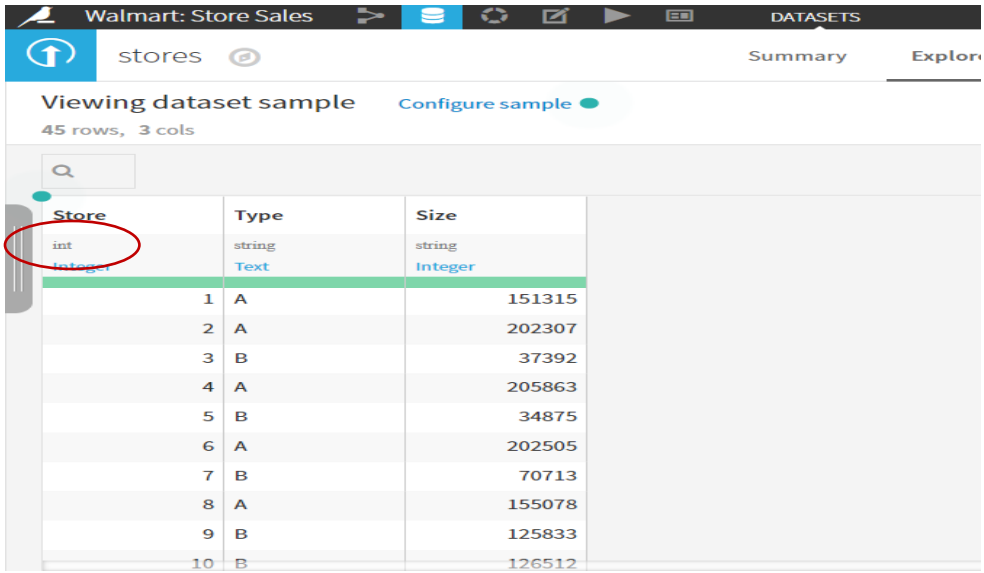
Store	Dept	Date	IsHoliday
string	string	string	string
Integer	Integer	Date (unparsed)	Boolean
1	1	2012-11-02	FALSE
1	1	2012-11-09	FALSE
1	1	2012-11-16	FALSE
1	1	2012-11-23	TRUE
1	1	2012-11-30	FALSE
1	1	2012-12-07	FALSE
1	1	2012-12-14	FALSE
1	1	2012-12-21	FALSE
1	1	2012-12-28	TRUE
1	1	2013-01-04	FALSE

STORE DATASET

In store dataset, I just change the storage type of "Store" from **string** to **int**. The reason why I changed it because later on, I will join it to feature and train dataset and feature and test dataset.



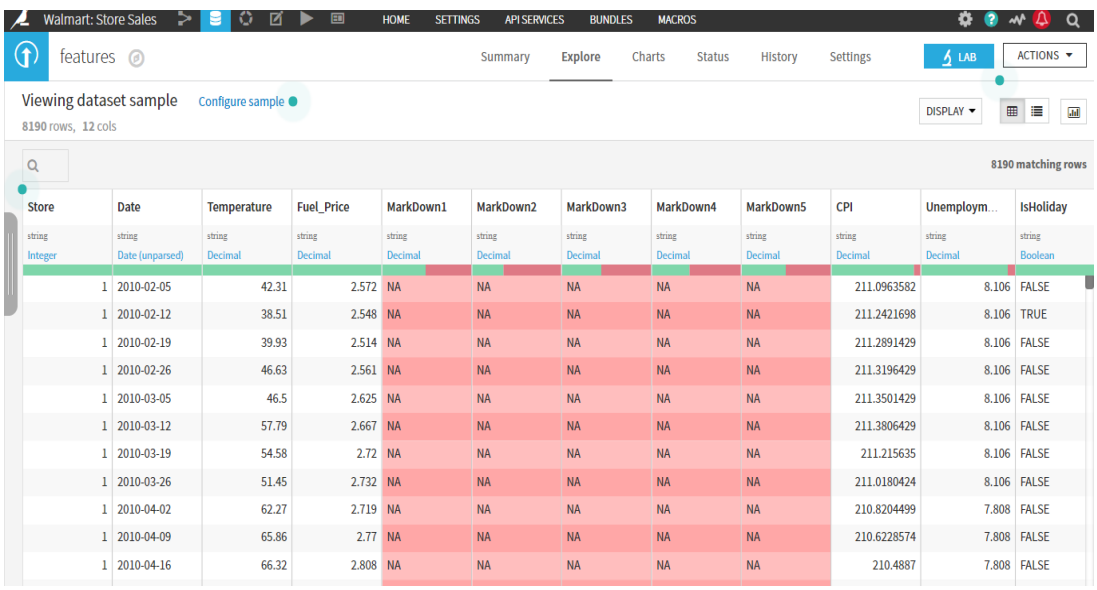
Store	Type	Size
string	string	string
Integer	Text	Integer
1	A	151315
2	A	202307
3	B	37392
4	A	205863
5	B	34875
6	A	202505
7	B	70713
8	A	155078
9	B	125833



Store	Type	Size
int Integer	string Text	string Integer
1	A	151315
2	A	202307
3	B	37392
4	A	205863
5	B	34875
6	A	202505
7	B	70713
8	A	155078
9	B	125833
10	B	126512

<u>FEATURE DATASET</u> (Using PREPARE RECIPE)

In this dataset, I did a lot of changes but not too much. As we all know, the meanings are automatically detected from the contents of the columns. In Markdown 1-5, CPI, and Unemployment, there should be done some transformation.

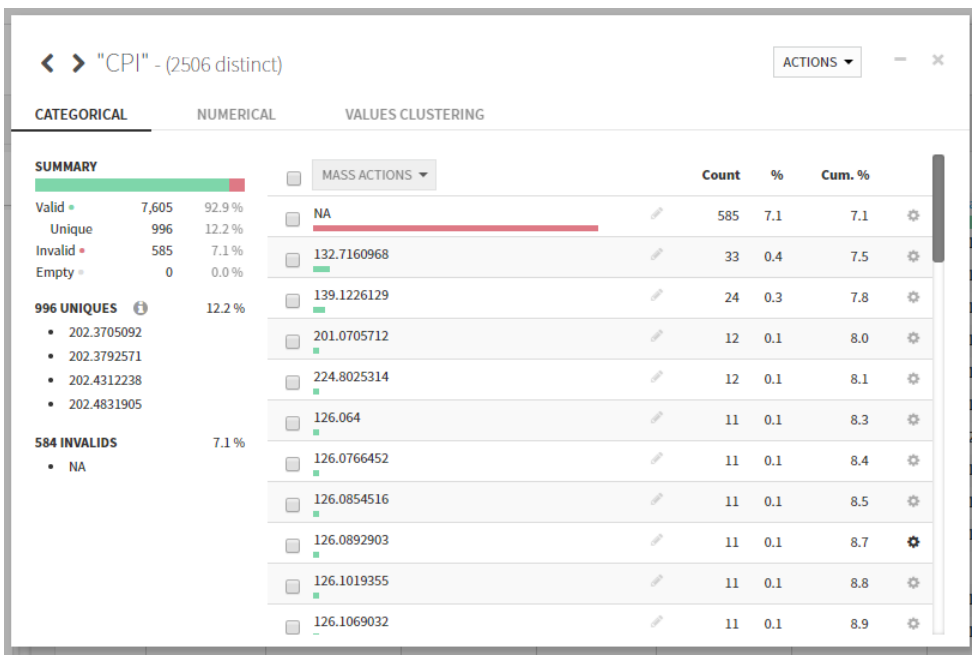


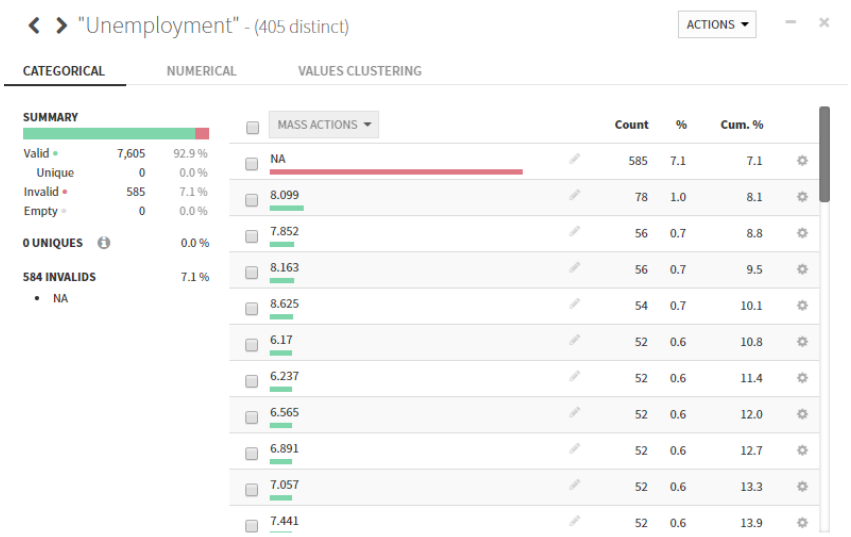
Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemploym...	IsHoliday
string Integer	string Date (unparsed)	string Decimal	string Decimal	string Decimal	string Decimal	string Decimal	string Decimal	string Decimal	string Decimal	string Decimal	string Boolean
1	2010-02-05	42.31	2.572	NA	NA	NA	NA	NA	211.0963582	8.106	FALSE
1	2010-02-12	38.51	2.548	NA	NA	NA	NA	NA	211.2421698	8.106	TRUE
1	2010-02-19	39.93	2.514	NA	NA	NA	NA	NA	211.2891429	8.106	FALSE
1	2010-02-26	46.63	2.561	NA	NA	NA	NA	NA	211.3196429	8.106	FALSE
1	2010-03-05	46.5	2.625	NA	NA	NA	NA	NA	211.3501429	8.106	FALSE
1	2010-03-12	57.79	2.667	NA	NA	NA	NA	NA	211.3806429	8.106	FALSE
1	2010-03-19	54.58	2.72	NA	NA	NA	NA	NA	211.215635	8.106	FALSE
1	2010-03-26	51.45	2.732	NA	NA	NA	NA	NA	211.0180424	8.106	FALSE
1	2010-04-02	62.27	2.719	NA	NA	NA	NA	NA	210.8204499	7.808	FALSE
1	2010-04-09	65.86	2.77	NA	NA	NA	NA	NA	210.6228574	7.808	FALSE
1	2010-04-16	66.32	2.808	NA	NA	NA	NA	NA	210.4887	7.808	FALSE

First, I need to change the NA values of Markdown 1-5. I replaced the NA values with another value called -999999. I put something different from zero because I saw that zero is a possible valid value. Look the result below:

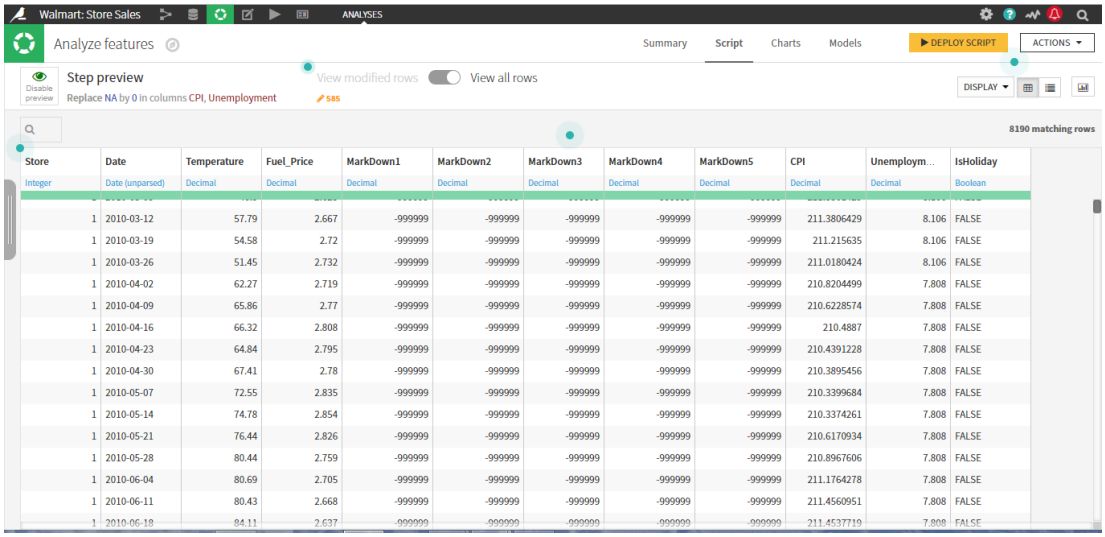
MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5
Decimal	Decimal	Decimal	Decimal	Decimal
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999
-999999	-999999	-999999	-999999	-999999

Second, the CPI and Unemployment have also NA values (look the pictures below). And, I don't want to delete those rows because there is a possibility that it will affect the final dataset later on. So, I decided to put 0 as their values.



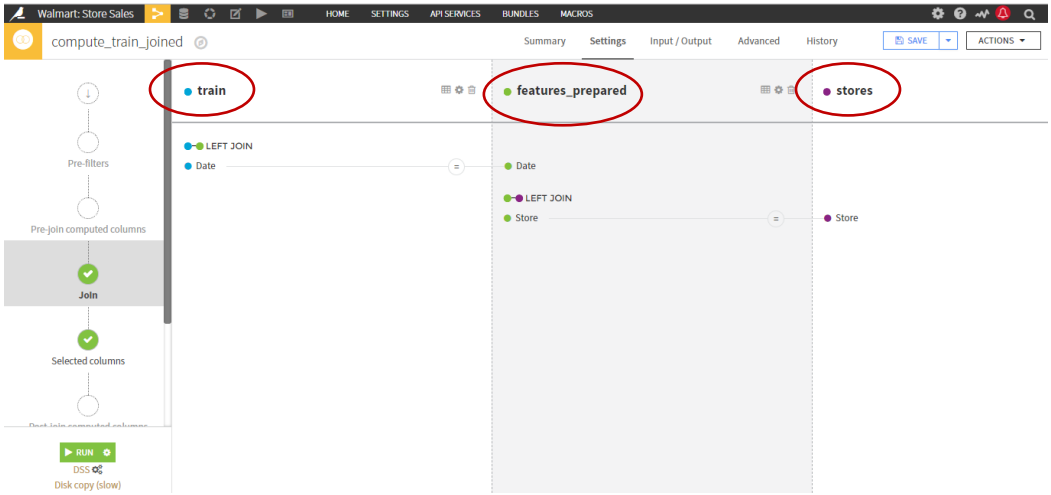


So, here is the cleaned feature dataset.

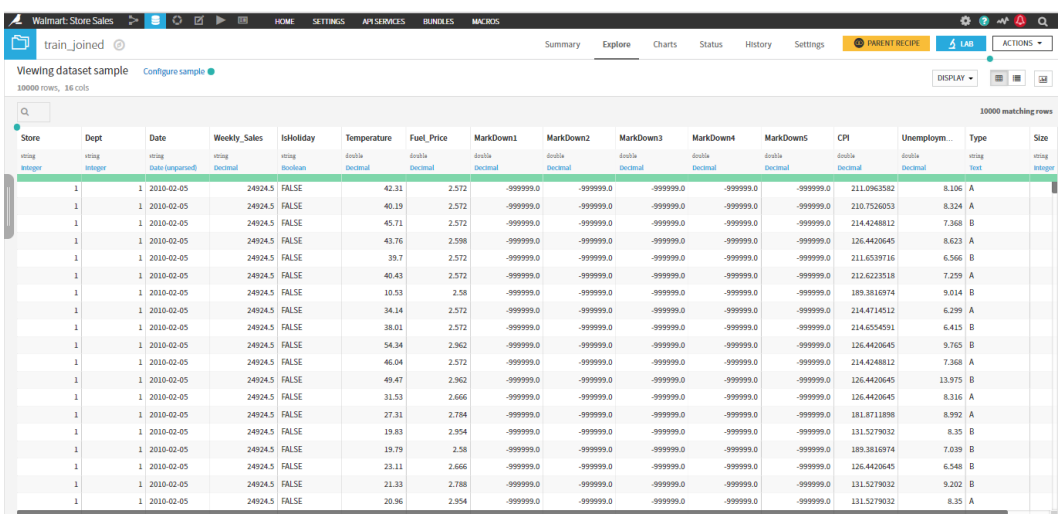


- Let's go to **JOIN RECIPE** which is dedicated to joins between two or more datasets. I merged the three data files (train, features and stores) to see the effect of different variables on sale. I merged also the test, features and stores dataset. Let's have a look.

TRAIN – FEATURES_PREPARED - STORES

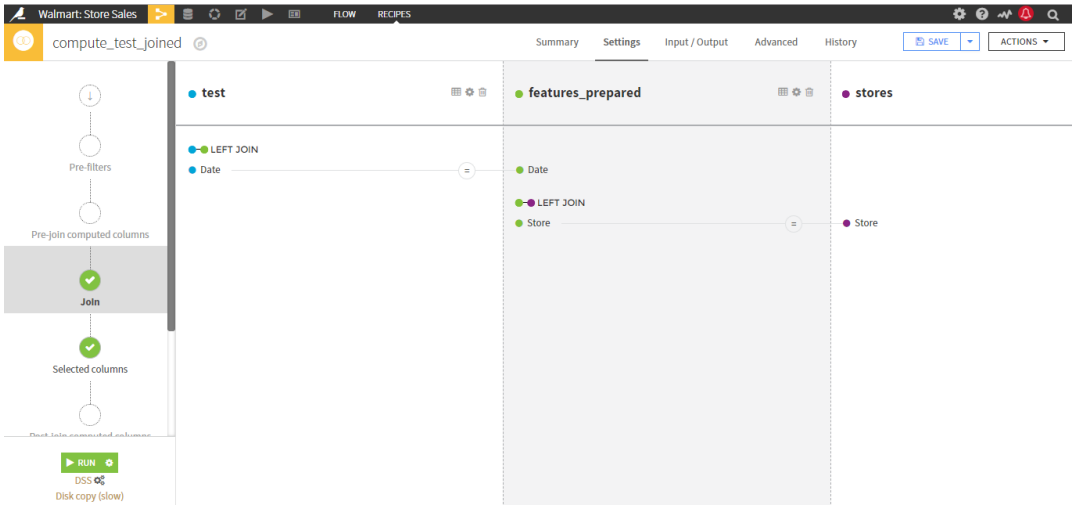


TRAIN_JOINED

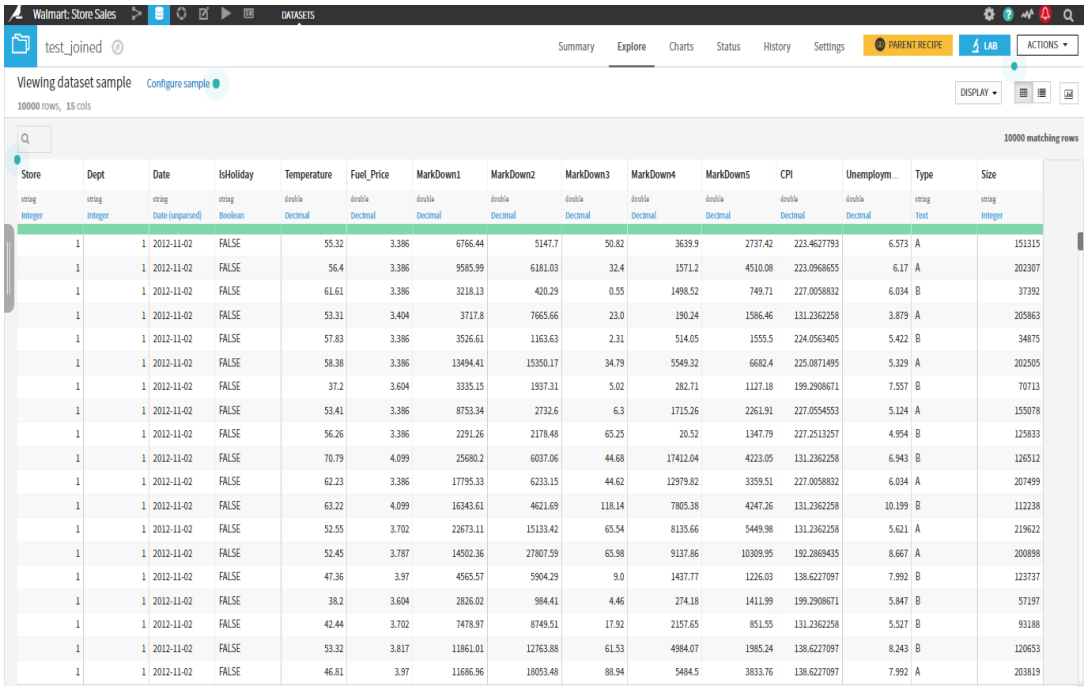


Store	Dept	Date	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemploym...	Type	Size
string	string	string	string	string	double	double	double	double	double	double	double	double	double	string	string
Integer	Integer	Date (unparsed)	Decimal	Boolean	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Text	Integer
1	1	2010-02-05	24924.5	FALSE	42.31	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	211.0963582	8.106	A	
1	1	2010-02-05	24924.5	FALSE	40.19	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	210.7526053	8.324	A	
1	1	2010-02-05	24924.5	FALSE	45.71	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	214.4248812	7.368	B	
1	1	2010-02-05	24924.5	FALSE	43.76	2.598	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	126.4420645	8.623	A	
1	1	2010-02-05	24924.5	FALSE	39.7	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	211.6539716	6.566	B	
1	1	2010-02-05	24924.5	FALSE	40.43	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	212.6223518	7.259	A	
1	1	2010-02-05	24924.5	FALSE	10.53	2.58	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	189.3816974	9.014	B	
1	1	2010-02-05	24924.5	FALSE	34.14	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	214.4714512	6.299	A	
1	1	2010-02-05	24924.5	FALSE	38.01	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	214.6554591	6.415	B	
1	1	2010-02-05	24924.5	FALSE	54.34	2.962	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	126.4420645	9.765	B	
1	1	2010-02-05	24924.5	FALSE	46.04	2.572	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	214.4248812	7.368	A	
1	1	2010-02-05	24924.5	FALSE	49.47	2.962	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	126.4420645	13.975	B	
1	1	2010-02-05	24924.5	FALSE	31.53	2.666	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	126.4420645	8.316	A	
1	1	2010-02-05	24924.5	FALSE	27.31	2.784	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	181.8711898	8.992	A	
1	1	2010-02-05	24924.5	FALSE	19.83	2.954	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	131.5279032	8.35	B	
1	1	2010-02-05	24924.5	FALSE	19.79	2.58	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	189.3816974	7.039	B	
1	1	2010-02-05	24924.5	FALSE	23.11	2.666	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	126.4420645	6.548	B	
1	1	2010-02-05	24924.5	FALSE	21.33	2.788	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	131.5279032	9.202	B	
1	1	2010-02-05	24924.5	FALSE	20.96	2.954	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	131.5279032	8.35	A	

TEST – FEATURES_PREPARED – STORES



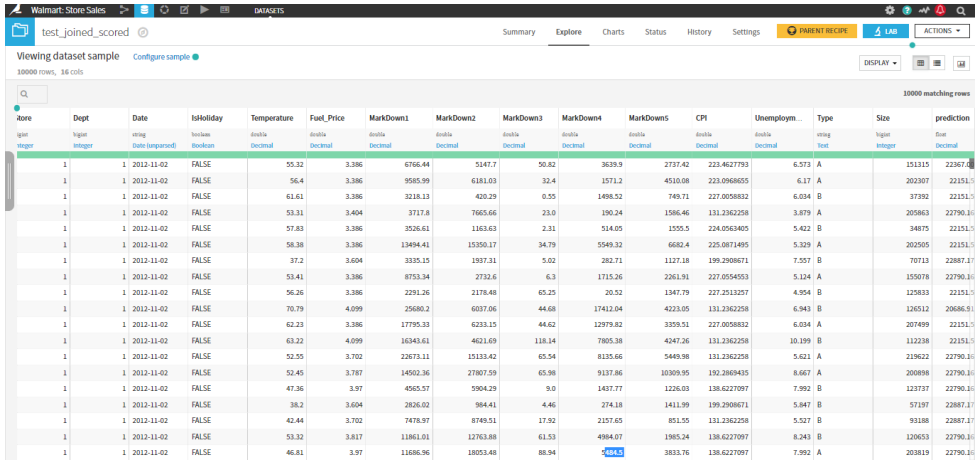
TEST_JOINED



Store	Dept	Date	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemploym...	Type	Size
string	string	string	string	double	double	double	double	double	double	double	double	double	string	string
Integer	Integer	Date (unparsed)	Boolean	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Text	Integer
1	1	2012-11-02	FALSE	55.32	3.386	6766.44	5147.7	50.82	3639.9	2737.42	223.4627793	6.573	A	151315
1	1	2012-11-02	FALSE	56.4	3.386	9585.99	6181.03	32.4	1571.2	4510.08	223.0968655	6.17	A	202307
1	1	2012-11-02	FALSE	61.61	3.386	3218.13	420.29	0.55	1498.52	749.71	227.0058832	6.034	B	37392
1	1	2012-11-02	FALSE	53.31	3.404	3717.8	7665.66	23.0	190.24	1586.46	131.2362258	3.879	A	205863
1	1	2012-11-02	FALSE	57.83	3.386	3526.61	1163.63	2.31	514.05	1555.5	224.0563405	5.422	B	34875
1	1	2012-11-02	FALSE	58.38	3.386	13494.41	15350.17	34.79	5549.32	6682.4	225.0871495	5.329	A	202505
1	1	2012-11-02	FALSE	37.2	3.604	3335.15	1937.31	5.02	282.71	1127.18	199.2908671	7.557	B	70713
1	1	2012-11-02	FALSE	53.41	3.386	8753.34	2732.6	6.3	1715.26	2261.91	227.0554553	5.124	A	155078
1	1	2012-11-02	FALSE	56.26	3.386	2291.26	2178.48	65.25	20.52	1347.79	227.2513257	4.954	B	125833
1	1	2012-11-02	FALSE	70.79	4.099	25680.2	6037.06	44.68	17412.04	4223.05	131.2362258	6.943	B	126512
1	1	2012-11-02	FALSE	62.23	3.386	17795.33	6233.15	44.62	12979.82	3359.51	227.0058832	6.034	A	207499
1	1	2012-11-02	FALSE	63.22	4.099	16343.61	4621.69	118.14	7805.38	4247.26	131.2362258	10.199	B	112238
1	1	2012-11-02	FALSE	52.55	3.702	22673.11	15133.42	65.54	8135.66	5449.98	131.2362258	5.621	A	219622
1	1	2012-11-02	FALSE	52.45	3.787	14502.36	27807.59	65.98	9137.86	10309.95	192.2869435	8.667	A	200898
1	1	2012-11-02	FALSE	47.36	3.97	4565.57	5904.29	9.0	1437.77	1226.03	138.6227097	7.992	B	123737
1	1	2012-11-02	FALSE	38.2	3.604	2826.02	984.41	4.46	274.18	1411.99	199.2908671	5.847	B	57197
1	1	2012-11-02	FALSE	42.44	3.702	7478.97	8749.51	17.92	2157.65	851.55	131.2362258	5.527	B	93188
1	1	2012-11-02	FALSE	53.32	3.817	11861.01	12763.88	61.53	4984.07	1985.24	138.6227097	8.243	B	120653
1	1	2012-11-02	FALSE	46.81	3.97	11686.96	18053.48	88.94	5484.5	3833.76	138.6227097	7.992	A	203819

- After I scored the model that I chose, I had a new dataset which I need to make some transformation to the dataset.

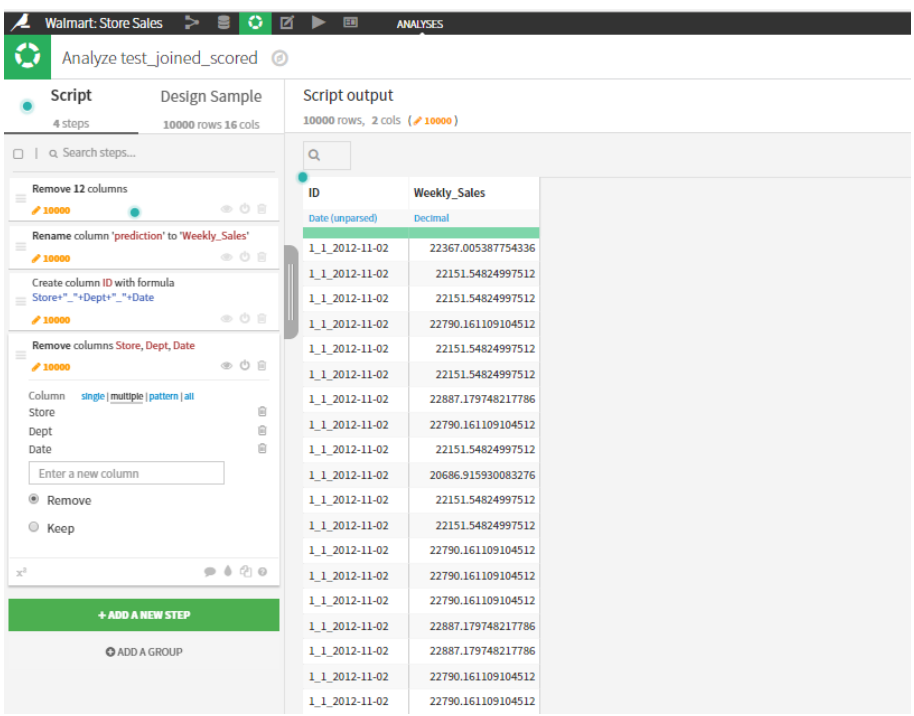
Test_joined_scored



Store	Dept	Date	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemploym...	Type	Size	prediction
bigint	bigint	string	boolean	double	double	double	double	double	double	double	double	double	string	bigint	float
integer	integer	Date (unparsed)	Boolean	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Text	Integer	Decimal
1	1	2012-11-02	FALSE	55.32	3.386	6766.44	5147.7	50.82	3639.9	2737.42	223.4627793	6.573	A	151315	22367.0
1	1	2012-11-02	FALSE	56.4	3.386	9585.99	6181.03	32.4	1571.2	4510.08	223.0968655	6.17	A	202307	22151.5
1	1	2012-11-02	FALSE	61.61	3.386	3218.13	420.29	0.55	1498.52	749.71	227.0058832	6.034	B	37392	22151.5
1	1	2012-11-02	FALSE	53.31	3.404	3717.8	7665.66	23.0	190.24	1586.46	131.2362258	3.879	A	205863	22790.16
1	1	2012-11-02	FALSE	57.83	3.386	3526.61	1163.63	2.31	514.05	1555.5	224.0563405	5.422	B	34875	22151.5
1	1	2012-11-02	FALSE	58.38	3.386	13494.41	15350.17	34.79	5549.32	6682.4	225.0871495	5.329	A	202505	22151.5
1	1	2012-11-02	FALSE	37.2	3.604	3335.15	1937.31	5.02	282.71	1127.18	199.2908671	7.557	B	70713	22887.17
1	1	2012-11-02	FALSE	53.41	3.386	8753.34	2732.6	6.3	1715.26	2261.91	227.0554553	5.124	A	155078	22790.16
1	1	2012-11-02	FALSE	56.26	3.386	2291.26	2178.48	65.25	20.52	1347.79	227.2513257	4.954	B	125833	22151.5
1	1	2012-11-02	FALSE	70.79	4.099	25680.2	6037.06	44.68	17412.04	4223.05	131.2362258	6.943	B	126512	20686.91
1	1	2012-11-02	FALSE	62.23	3.386	17795.33	6233.15	44.62	12979.82	3359.51	227.0058832	6.034	A	207499	22151.5
1	1	2012-11-02	FALSE	63.22	4.099	16343.61	4621.69	118.14	7805.38	4247.26	131.2362258	10.199	B	112238	22151.5
1	1	2012-11-02	FALSE	52.55	3.702	22673.11	15133.42	65.54	8135.66	5449.98	131.2362258	5.621	A	219622	22790.16
1	1	2012-11-02	FALSE	52.45	3.787	14502.36	27807.59	65.98	9137.86	10309.95	192.2869435	8.667	A	200898	22790.16
1	1	2012-11-02	FALSE	47.36	3.97	4565.57	5904.29	9.0	1437.77	1226.03	138.6227097	7.992	B	123737	22790.16
1	1	2012-11-02	FALSE	38.2	3.604	2826.02	984.41	4.46	274.18	1411.99	199.2908671	5.847	B	57197	22887.17
1	1	2012-11-02	FALSE	42.44	3.702	7478.97	8749.51	17.92	2157.65	851.55	131.2362258	5.527	B	93188	22887.17
1	1	2012-11-02	FALSE	53.32	3.817	11861.01	12763.88	61.53	4984.07	1985.24	138.6227097	8.243	B	120653	22790.16
1	1	2012-11-02	FALSE	46.81	3.97	11686.96	18053.48	88.94	4854.5	3833.76	138.6227097	7.992	A	203819	22790.16

Final Column of the Final Dataset

- I removed 12 columns which are IsHoliday, temperature, fuel_price, markdown 1-5, CPI, Unemployment, type and size
- I renamed the prediction column to weekly_sales
- I created a formula to concatenate the three columns



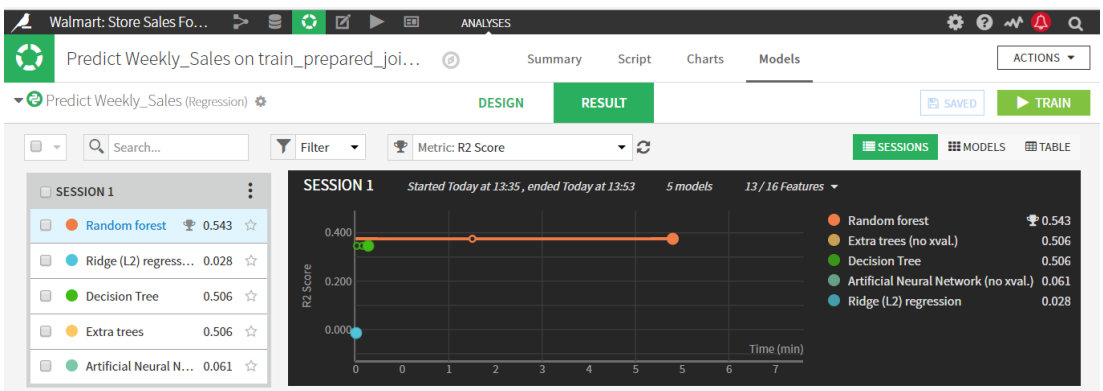
ID	Weekly_Sales
Date (unparsed)	Decimal
1_1_2012-11-02	22367.005387754336
1_1_2012-11-02	22151.54824997512
1_1_2012-11-02	22151.54824997512
1_1_2012-11-02	22790.161109104512
1_1_2012-11-02	22151.54824997512
1_1_2012-11-02	22151.54824997512
1_1_2012-11-02	22887.179748217786
1_1_2012-11-02	22790.161109104512
1_1_2012-11-02	22151.54824997512
1_1_2012-11-02	20686.915930083276
1_1_2012-11-02	22151.54824997512
1_1_2012-11-02	22151.54824997512
1_1_2012-11-02	22790.161109104512
1_1_2012-11-02	22790.161109104512
1_1_2012-11-02	22790.161109104512
1_1_2012-11-02	22887.179748217786
1_1_2012-11-02	22887.179748217786
1_1_2012-11-02	22790.161109104512
1_1_2012-11-02	22790.161109104512

Predictive Modeling

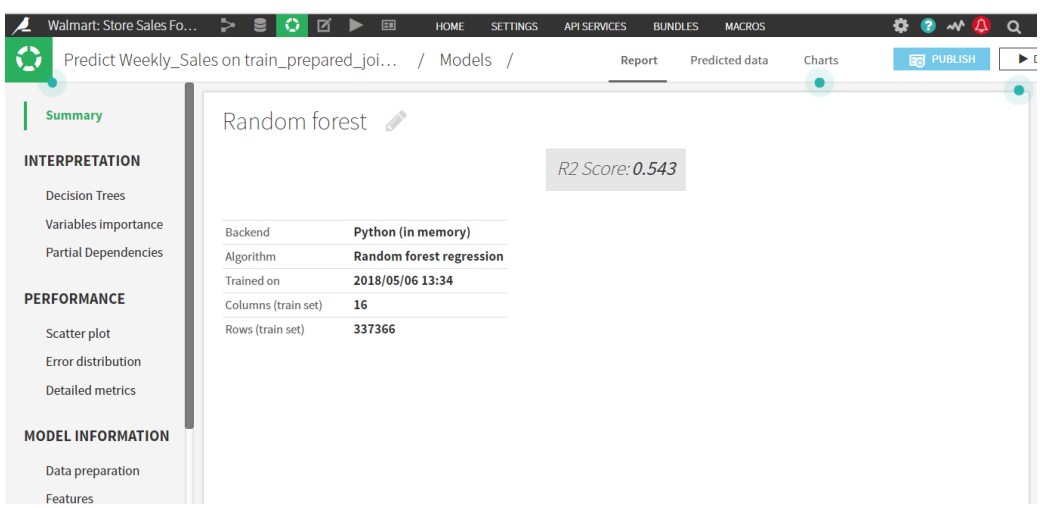
Building and deploying a predictive model. In this part, I can use a lot of model like random forest, decision tree, extra trees, artificial neutral network etc., then, at last I will choose the best model.

Now, I'm done to clean and do some engineering features in my dataset. So, it's time to do a prediction model. I used "weekly_sales" as my target when I did the prediction. I did a lot of models to train my dataset like Random Forest, Ridge (L2) regression, Decision Tree, Extra trees and artificial Neural Network. In my first flow that I had done, the Random Forest was in the top or recommended to use but then, when I did another flow for my project, the recommended one was the Extra trees. At last, I decided to use the Extra trees as my model. The reason why I did two different flow because I'm not satisfied from the result of my first flow. When I did, another session from my first flow, the result are always the same. Until, I decided to make another flow for my project.

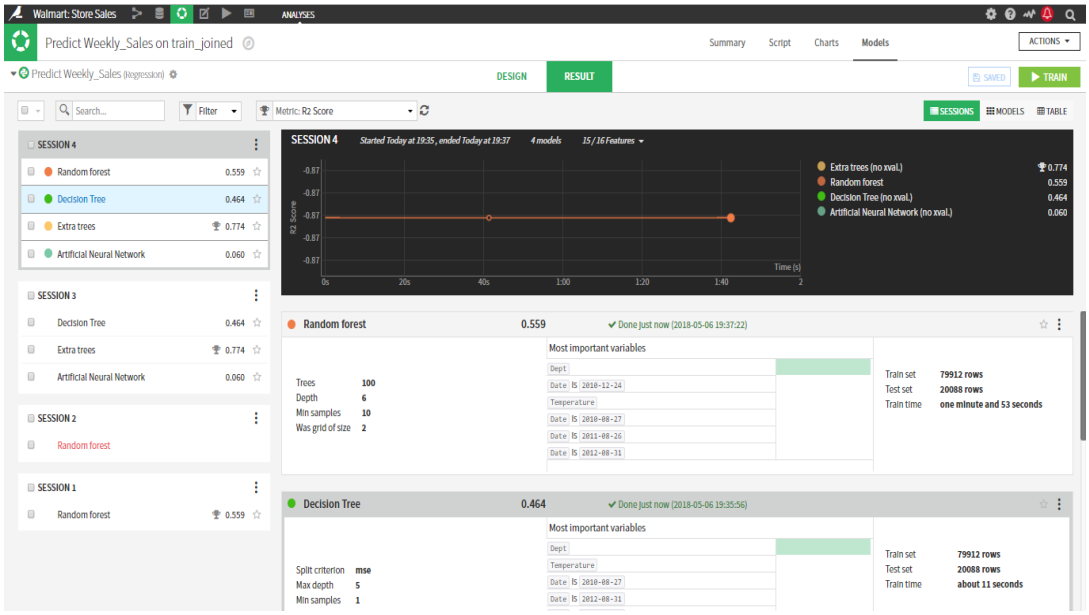
- **Below was the result of the first FLOW of my project that I had done and I chose Random Forest as the model.**

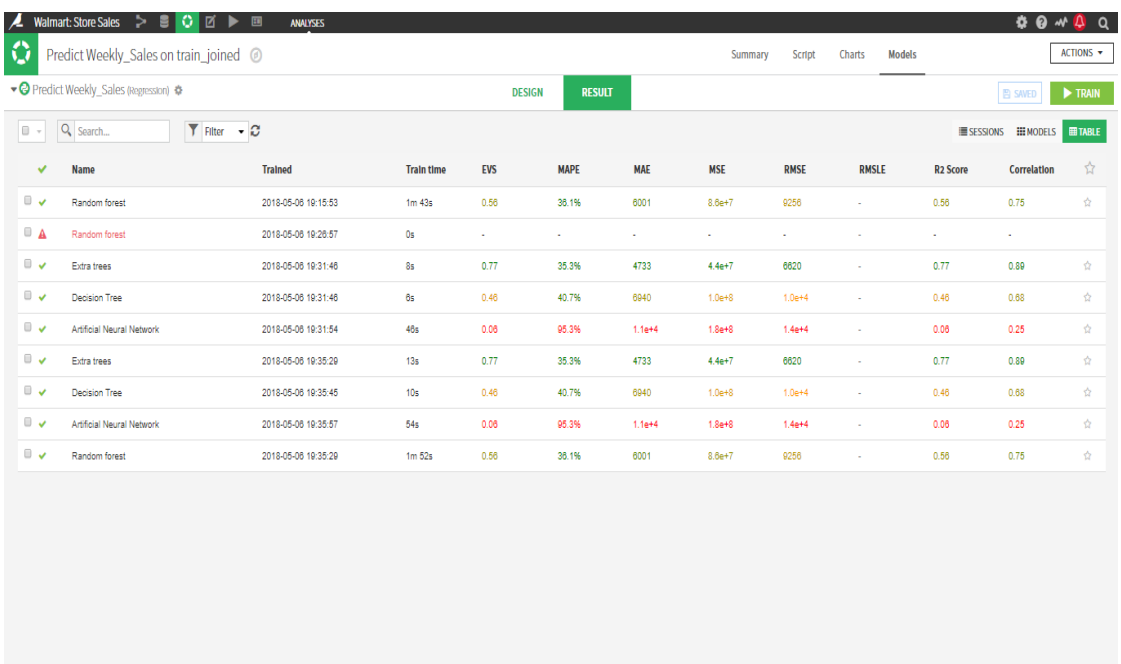


<u>RESULT</u>



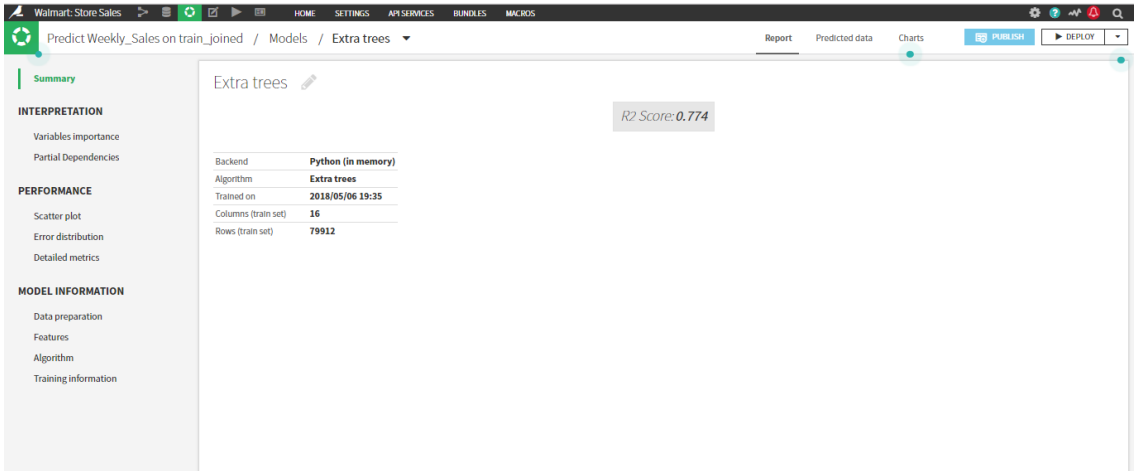
- **Below was the result of the second FLOW of my project that I had done and I chose Extra trees as the model.**





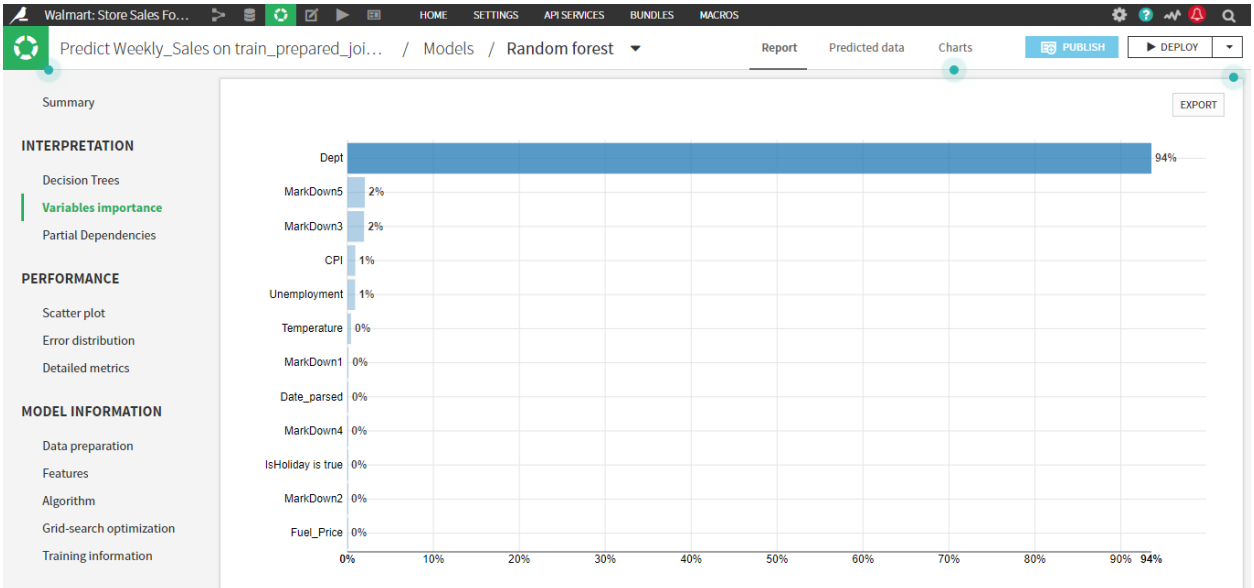
Name	Trained	Train time	EVS	MAPE	MAE	MSE	RMSE	RMSLE	R2 Score	Correlation
Random forest	2018-05-06 19:15:53	1m 43s	0.56	36.1%	6001	8.6e+7	9256	-	0.56	0.75
Random forest	2018-05-06 19:28:57	0s	-	-	-	-	-	-	-	-
Extra trees	2018-05-06 19:31:46	8s	0.77	35.3%	4733	4.4e+7	6620	-	0.77	0.89
Decision Tree	2018-05-06 19:31:46	6s	0.46	40.7%	6940	1.0e+8	1.0e+4	-	0.46	0.68
Artificial Neural Network	2018-05-06 19:31:54	46s	0.06	95.3%	1.1e+4	1.8e+8	1.4e+4	-	0.06	0.25
Extra trees	2018-05-06 19:35:29	13s	0.77	35.3%	4733	4.4e+7	6620	-	0.77	0.89
Decision Tree	2018-05-06 19:35:45	10s	0.46	40.7%	6940	1.0e+8	1.0e+4	-	0.46	0.68
Artificial Neural Network	2018-05-06 19:35:57	54s	0.06	95.3%	1.1e+4	1.8e+8	1.4e+4	-	0.06	0.25
Random forest	2018-05-06 19:35:29	1m 52s	0.56	36.1%	6001	8.6e+7	9256	-	0.56	0.75

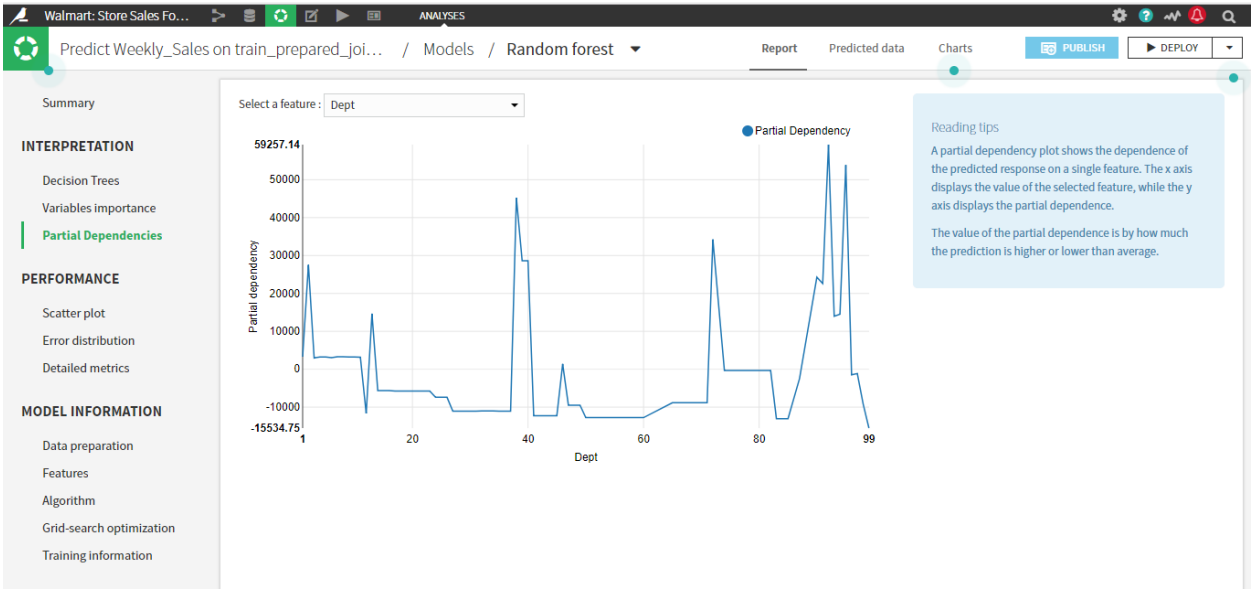
RESULT



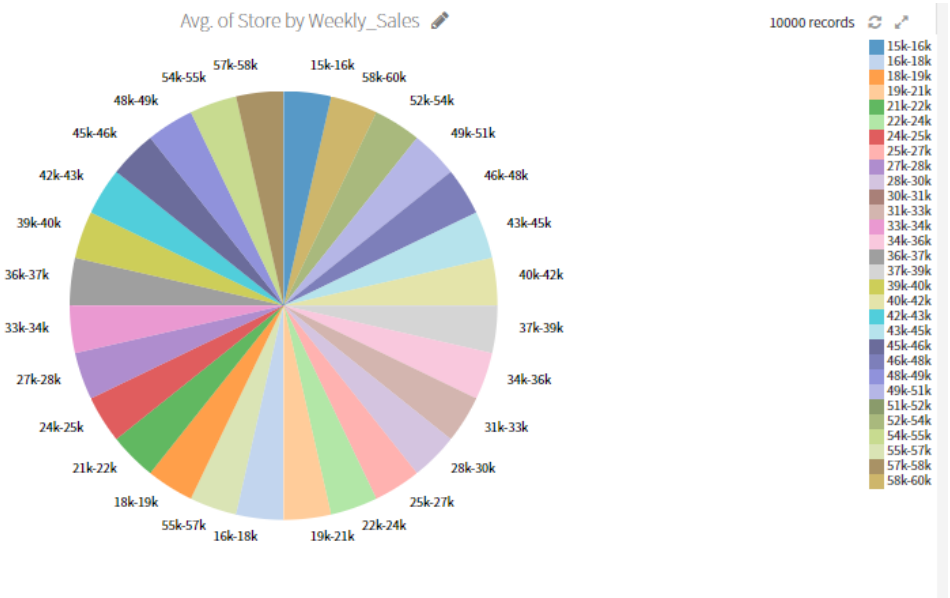
Visualization

Creating a useful visualization of our predicted data and the dataset.



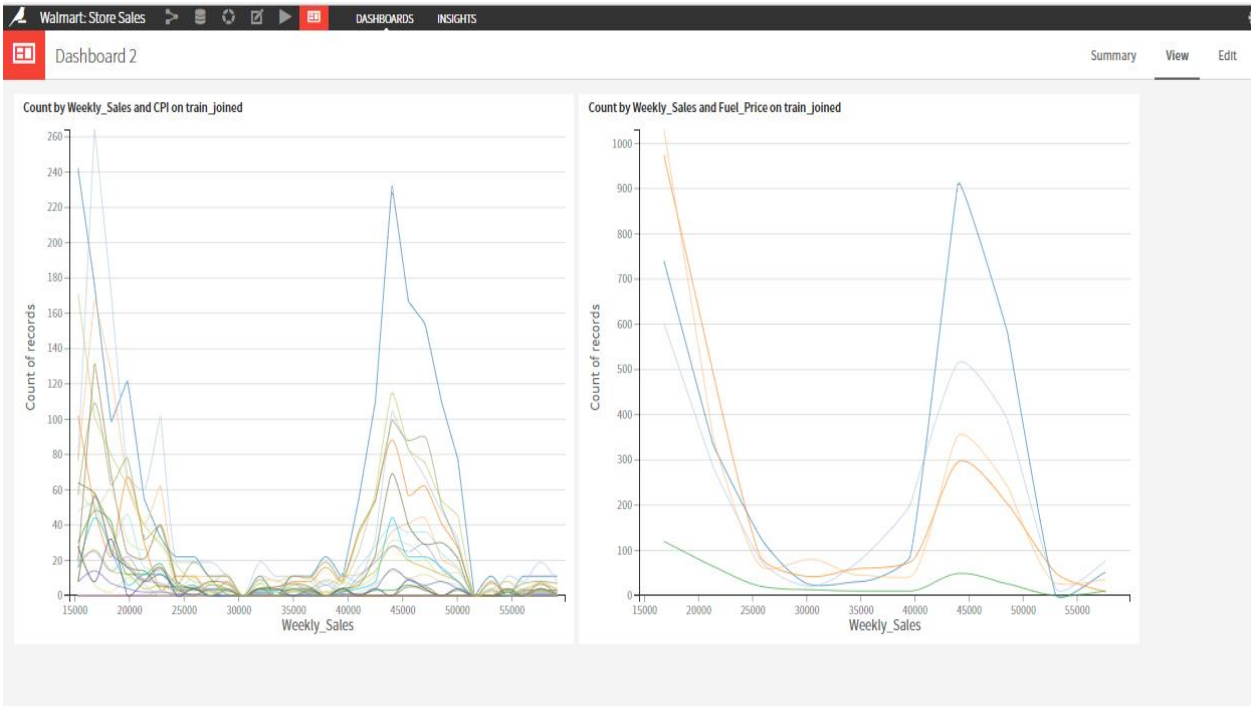


Weekly sales per store



DASHBOARDS

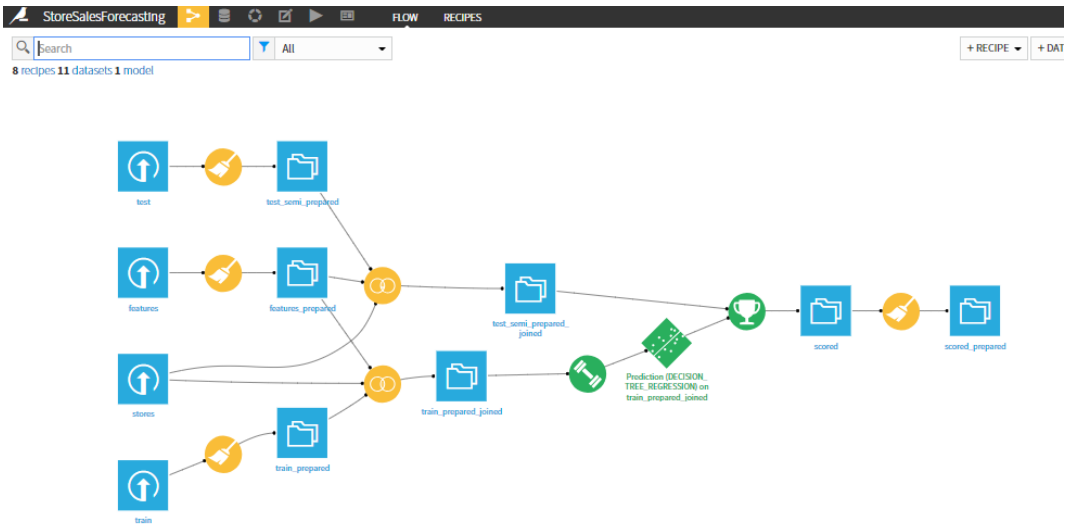
Walmart: Store Sales
DASHBOARDS
INSIGHTS
Default dashboard
Summary
View
Edit
ACTIONS
Avg. of Store by Weekly_Sales on train_joined
Avg. of Weekly_Sales by Fuel_Price and CPI on train_joined
Weekly_Sales (AVG)
Fuel_Price
Count by Weekly_Sales and Size on train_joined
Count of records
Weekly_Sales

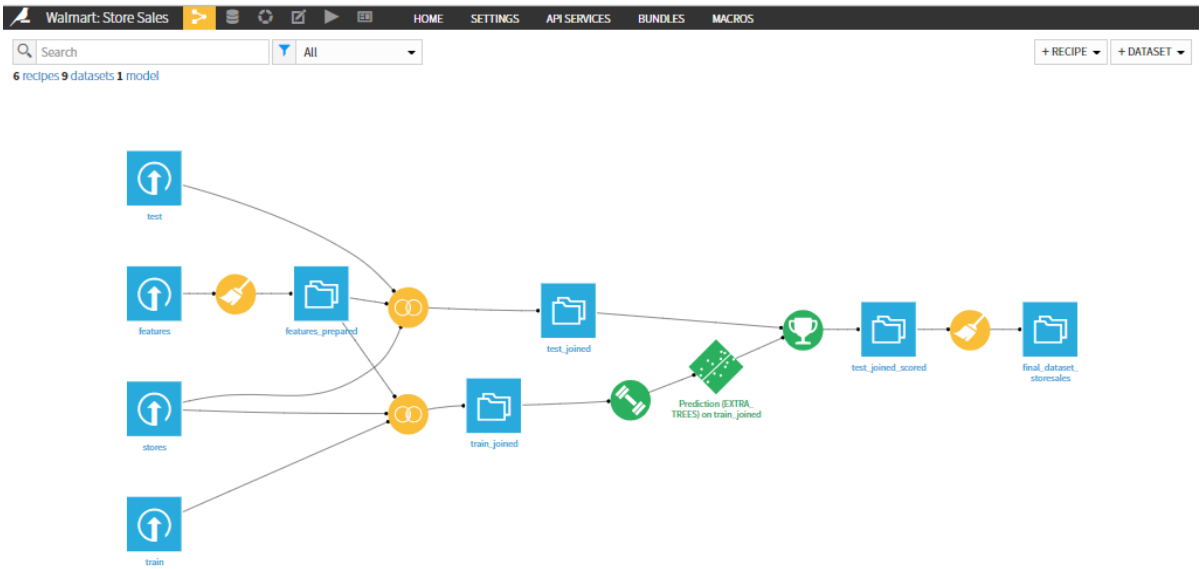

Walmart: Store Sales
DASHBOARDS
INSIGHTS
Dashboard 2
Summary
View
Edit
Count by Weekly_Sales and CPI on train_joined
Count of records
Weekly_Sales
Count by Weekly_Sales and Fuel_Price on train_joined
Count of records
Weekly_Sales

FLOW OF THE PROJECT

Actually, they are almost the same. At the end, I decided not to clean the test and train dataset (look at cleaning section the overview of train and test dataset).

First Flow of my Project



Second Flow of my Project

(This is my final flow)