DECEMBER 2022

MACHINE LEARNING

SUPERVISED MODELS

GROUP 5

ADRIAN KU
IGNACIO PIRE RUBIO
IVAN LOPEZ CANSADO
LAKSHMI PRIYANKA JAKKA
LIA DOLLISON
ROMAN ZOTKIN

1.EDA & DATA CLEANING

The objective of this part is to perform some Exploratory Data Analysis, EDA, clean data, and data pre-processing steps on this dataset.

1. Dataset insights
2. EDA
   - Descriptive Analysis
   - Data Cleaning /pre-processing
3. Final Dataset

2.MODELLING

Train a Machine Learning model to try to predict solar energy production of 3 stations (ACME, GOOD, and WYNO) using the given dataset.

1. ACME Model
2. GOOD Model
3. WYNO Model
4. Final conclusions/ouputs

1.EDA & DATA CLEANING

DATASET INSIGHTS

First of all, it is important to understand what we have, what are our goals, and how are we going to do it.

WHAT DO WE HAVE

For the execution of this assignment, we have 3 datasets:

solar_dataset.csv: **actual solar production values of different stations** from 01/01/1994 to 31/12/2007 (from 01/01/2008 values are missing).
This dataset has 456 columns and 5370 rows.
The first 99 columns (except for the first one- date), indicate the real values of solar production of each of the stations per day.
The remaining variables (from the 100th column to 456th) are PCAs (Principal Components Analysis) over the original dataset, given by weather predictors.
From row 5113 (01/01/2008), there are NA or missing values.

additional_variables.csv: **Real Numerical Weather Prediction**, NWP. The 100 most important variables to predict the ACME station (One of our target variables to predict). Column 1 = Date

station_info.csv: **Information** about the original dataset. File with name, latitude, longitude, and elevation of each of the 98 stations. For Exploratory Data Analysis, **EDA**.

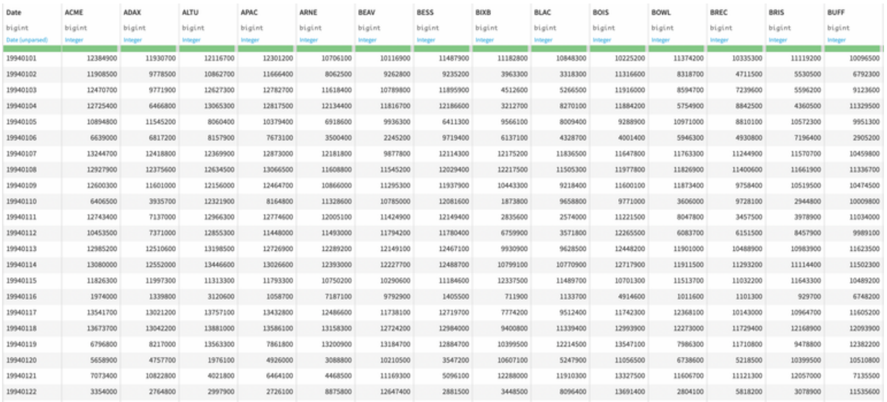
So, we will treat with solar_dataset.csv and additional_variables.csv.

GOAL
Prediction of the solar production of stations ACME, GOOD, and WYNO from 01/01/2008 to 30/11/2012

HOW
Machine Learning supervised models (Regression), using Dataiku.

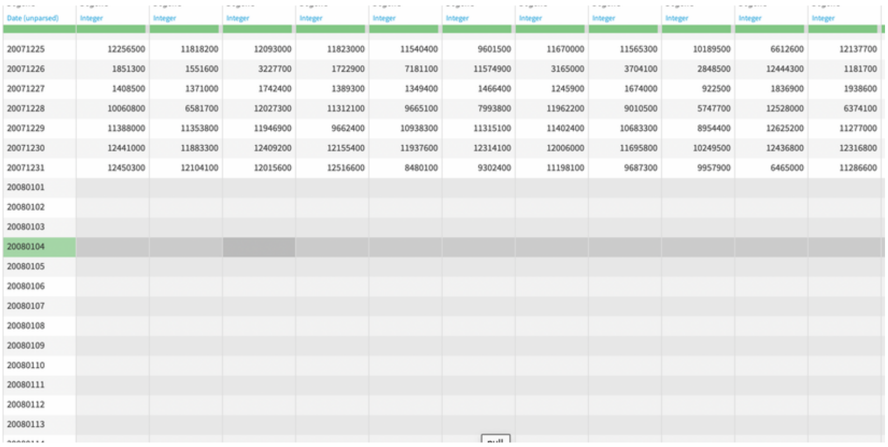
1.EDA & DATA CLEANING

solar production per station



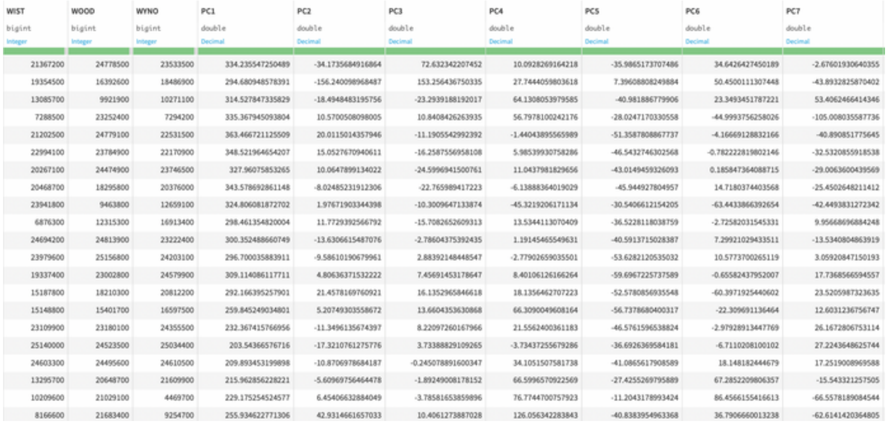
Date	ACME	ADAX	ALTU	APAC	ARNE	BEAV	BESS	BIXB	BLAC	BOIS	BOWL	BREC	BRIS	BUFF
bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint
Date (unparsed)	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer
19940101	12384900	11930700	12116700	12301200	10706100	10116900	11487900	11182800	10848300	10225200	11374200	10335300	11119200	10096500
19940102	11908500	9778500	10862700	11666400	8062500	9262800	9235200	3963300	3318300	11316600	8318700	4711500	5530500	6792300
19940103	12470700	9771900	12627300	12782700	11618400	10789800	11895900	4512600	5266500	11916000	8594700	7239600	5596200	9123600
19940104	12725400	6466800	13065300	12817500	12134400	11816700	12186600	3212700	8270100	11884200	5754900	8842500	4360500	11329500
19940105	10894800	11545200	8060400	10379400	6918600	9936300	6411300	9566100	8009400	9288900	10971000	8810100	10572300	9951300
19940106	6639000	6817200	8157900	7673100	3500400	2245200	9719400	6137100	4328700	4001400	5946300	4930800	7196400	2905200
19940107	13244700	12418800	12369900	12873000	12181800	9877800	12114300	12175200	11836500	11647800	11763300	11244900	11570700	10459800
19940108	12927900	12375600	12634500	13066500	11608800	11545200	12029400	12217500	11505300	11977800	11826900	11400600	11661900	11336700
19940109	12600300	11601000	12156000	12464700	10866000	11295300	11937900	10443300	9218400	11600100	11873400	9758400	10519500	10474500
19940110	6406500	3935700	12321900	8164800	11328600	10785000	12081600	1873800	9658800	9771000	3606000	9728100	2944800	10009800
19940111	12743400	7137000	12966300	12774600	12005100	11424900	12149400	2835600	2574000	11221500	8047800	3457500	3978900	11034000
19940112	10453500	7371000	12855300	11448000	11493000	11794200	11780400	6759900	3571800	12265500	6083700	6151500	8457900	9989100
19940113	12985200	12510600	13198500	12726900	12289200	12149100	12467100	9930900	9628500	12448200	11901000	10488900	10983900	11623500
19940114	13080000	12552000	13446600	13026600	12393000	12227700	12488700	10799100	10770900	12717900	11911500	11293200	11114400	11502300
19940115	11826300	11997300	11313300	11793300	10750200	10290600	11184600	12337500	11489700	10701300	11513700	11032200	11643300	10489200
19940116	1974000	1339800	3120600	1058700	7187100	9792900	1405500	711900	1133700	4914600	1011600	1101300	929700	6748200
19940117	13541700	13021200	13757100	13432800	12486600	11738100	12719700	7774200	9512400	11742300	12368100	10143000	10964700	11605200
19940118	13673700	13042200	13881000	13586100	13158300	12724200	12984000	9400800	11339400	12993900	12273000	11729400	12168900	12093900
19940119	6796800	8217000	13563300	7861800	13200900	13184700	12884700	10399500	12214500	13547100	7986300	11710800	9478800	12382200
19940120	5658900	4757700	1976100	4926000	3088800	10210500	3547200	10607100	5247900	11056500	6738600	5218500	10399500	10510800
19940121	7073400	10822800	4021800	6464100	4468500	11169300	5096100	12288000	11910300	13327500	11606700	11121300	12057000	7135500
19940122	3354000	2764800	2997900	2726100	8875800	12647400	2881500	3448500	8096400	13691400	2804100	5818200	3078900	11535600

missing values



Date (unparsed)	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer
20071225	12256500	11818200	12093000	11823000	11540400	9601500	11670000	11565300	10189500	6612600	12137700
20071226	1851300	1551600	3227700	1722900	7181100	11574900	3165000	3704100	2848500	12444300	1181700
20071227	1408500	1371000	1742400	1389300	1349400	1466400	1245900	1674000	922500	1836900	1938600
20071228	10060800	6581700	12027300	11312100	9665100	7993800	11962200	9010500	5747700	12528000	6374100
20071229	11388000	11353800	11946900	9662400	10938300	11315100	11402400	10683300	8954400	12625200	11277000
20071230	12441000	11883300	12409200	12155400	11937600	12314100	12006000	11695800	10249500	12436800	12316800
20071231	12450300	12104100	12015600	12516600	8480100	9302400	11198100	9687300	9957900	6465000	11286600
20080101											
20080102											
20080103											
20080104											
20080105											
20080106											
20080107											
20080108											
20080109											
20080110											
20080111											
20080112											
20080113											

PCAs



WIST	WOOD	WYNO	PC1	PC2	PC3	PC4	PC5	PC6	PC7
bigint	bigint	bigint	double	double	double	double	double	double	double
Integer	Integer	Integer	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal
21367200	24778500	23533500	334.235547250489	-34.1735684916864	72.632342207452	10.0928269164218	-35.9865173707486	34.6426427450189	-2.67601930640355
19354500	16392600	18486900	294.680948578391	-156.240098968487	153.256436750335	27.7444059803618	7.39608808249884	50.4500111307448	-43.8932825870402
13085700	9921900	10271100	314.527847335829	-18.4948483195756	-23.2939188192017	64.1308053979585	-40.981886779906	23.3493451787221	53.4062466414346
7288500	23252400	7294200	335.367945093804	10.5700508098005	10.8408426263935	56.7978100242176	-28.0247170330558	-44.9993756258026	-105.008035587736
21202500	24779100	22531500	363.466721125509	20.0115014357946	-11.1905542992392	-1.44043895565989	-51.3587808867737	-4.16669128832166	-40.890851775645
22994100	23784900	22170900	348.521964654207	15.0527670940611	-16.2587556958108	5.98539930758286	-46.5432746302568	-0.782222819802146	-32.5320855918538
20267100	24474900	23746500	327.96075853265	10.0647899134022	-24.5996941500761	11.0437981829656	-43.0149459326093	0.185847364088715	-29.0063600439569
20468700	18295800	20376000	343.578692861148	-8.02485231912306	-22.765989417223	-6.13888364019029	-45.944927804957	14.7180374403568	-25.4502648211412
23941800	9463800	12659100	324.806081872702	1.97671903344398	-10.3009647133874	-45.3219206171134	-30.5406612154205	-63.4433866392654	-42.4493831272342
6876300	12315300	16913400	298.461354820004	11.7729392566792	-15.7082652609313	13.5344113070409	-36.5228118038759	-2.72582031545331	9.95668696884248
24694200	24813900	23222400	300.352488660749	-13.6306615487076	-2.78604375392435	1.19145465549631	-40.5913715028387	7.29921029433511	-13.5340804863919
23979600	25156800	24203100	296.700035883911	-9.58610190679961	2.88392148448547	-2.77902659035501	-53.6282120535032	10.5773700265119	3.05920847150193
19337400	23002800	24579900	309.114086117711	4.80636371532222	7.45691453178647	8.40106126166264	-59.6967225737589	-0.65582437952007	17.7368566594557
15187800	18210300	20812200	292.166395257901	21.4578169760921	16.1352965846618	18.1356462707223	-52.5780856935548	-60.3971925440602	23.5205987323635
15148800	15401700	16597500	259.845249034801	5.20749303558672	13.6604353630868	66.3090049608164	-56.7378680400317	-22.309691136464	12.6031236756747
23109900	23180100	24355500	232.367415766956	-11.3496135674397	8.22097260167966	21.5562400361183	-46.5761596538824	-2.97928913447769	26.1672806753114
25140000	24523500	25034400	203.54366576716	-17.3210761275776	3.73388829109265	-3.73437255679286	-36.6926369584181	-6.7110208100102	27.2243648625744
24603300	24495600	24610500	209.893453199898	-10.8706978684187	-0.245078891600347	34.1051507581738	-41.0865617908589	18.148182444679	17.2519008969588
13295700	20648700	21609900	215.962856228221	-5.60969756464478	-1.89249008178152	66.5996570922569	-27.4255269795889	67.2852209806357	-15.543321257505
10209600	21029100	4469700	229.175254524577	6.45406632884049	-3.78581653859896	76.7744700757923	-11.2043178993424	86.4566155416613	-66.5578189084544
8166600	21683400	9254700	255.934622771306	42.9314661657033	10.4061273887028	126.056342283843	-40.8383954963368	36.7906660013238	-62.6141420364805

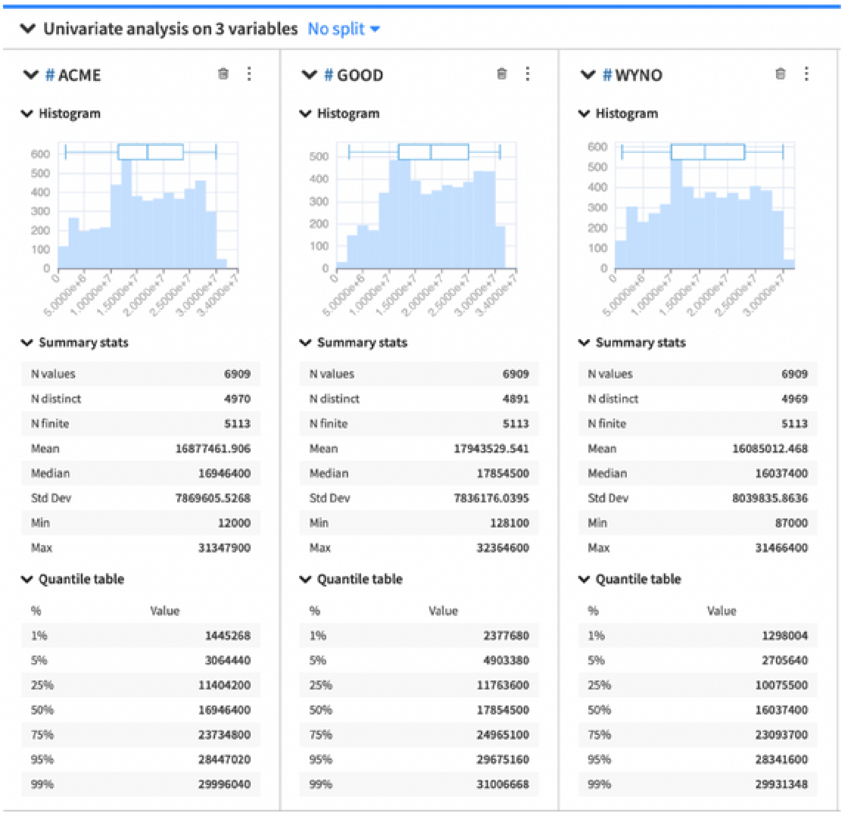
1.EDA & DATA CLEANING

EDA

Approach to analyzing datasets to summarize their main characteristics, often with visual methods. Two goals:
1. Seeing what the data can tell us before the modelling task. - **Descriptive Analysis**
2. First cleaning (or pre-processing) step before the modelling stage. - **Data cleaning /pre-processing**

1. DESCRIPTIVE ANALYSIS

a. solar_dataset.csv



ACME

Summary stats

N values	6909
N distinct	4970
N finite	5113
Mean	16877461.906
Median	16946400
Std Dev	7869605.5268
Min	12000
Max	31347900

Quantile table

%	Value
1%	1445268
5%	3064440
25%	11404200
50%	16946400
75%	23734800
95%	28447020
99%	29996040

GOOD

Summary stats

N values	6909
N distinct	4891
N finite	5113
Mean	17943529.541
Median	17854500
Std Dev	7836176.0395
Min	128100
Max	32364600

Quantile table

%	Value
1%	2377680
5%	4903380
25%	11763600
50%	17854500
75%	24965100
95%	29675160
99%	31006668

WYNO

Summary stats

N values	6909
N distinct	4969
N finite	5113
Mean	16085012.468
Median	16037400
Std Dev	8039835.8636
Min	87000
Max	31466400

Quantile table

%	Value
1%	1298004
5%	2705640
25%	10075500
50%	16037400
75%	23093700
95%	28341600
99%	29931348

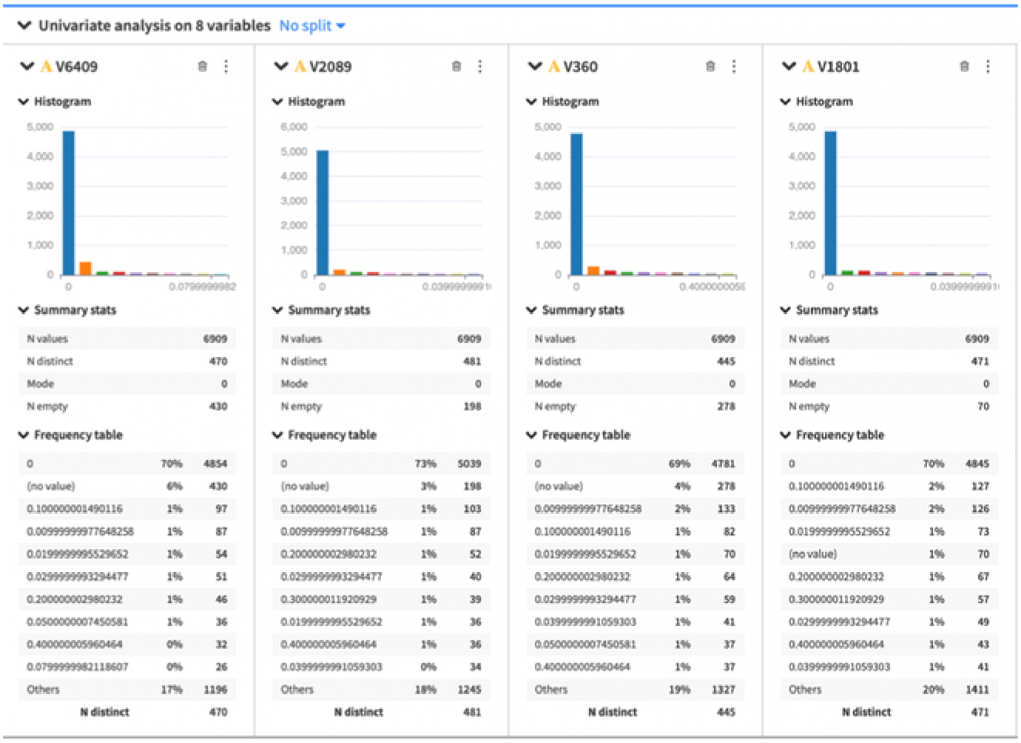
As we are just interested in the target variables: ACME, GOOD, and WYNO. We have done the descriptive analysis for these ones.

Conclusions:
- 99% of values are between [1298004 - 32364600].
- All missing values are from 01/01/2008 and onwards.
- Bigint data types.

1.EDA & DATA CLEANING

b. additional_variables.csv

As there are 100 variables, we have selected randomly 8 variables from them.



Univariate analysis on 8 variables No split

V6409

Summary stats

N values	6909
N distinct	470
Mode	0
N empty	430

Frequency table

0	70%	4854
(no value)	6%	430
0.100000001490116	1%	97
0.00999999977648258	1%	87
0.0199999995529652	1%	54
0.0299999993294477	1%	51
0.200000002980232	1%	46
0.0500000007450581	1%	36
0.400000005960464	0%	32
0.0799999982118607	0%	26
Others	17%	1196
N distinct		470

V2089

Summary stats

N values	6909
N distinct	481
Mode	0
N empty	198

Frequency table

0	73%	5039
(no value)	3%	198
0.100000001490116	1%	103
0.00999999977648258	1%	87
0.200000002980232	1%	52
0.0299999993294477	1%	40
0.300000011920929	1%	39
0.0199999995529652	1%	36
0.400000005960464	1%	36
0.0399999991059303	0%	34
Others	18%	1245
N distinct		481

V360

Summary stats

N values	6909
N distinct	445
Mode	0
N empty	278

Frequency table

0	69%	4781
(no value)	4%	278
0.00999999977648258	2%	133
0.100000001490116	1%	82
0.0199999995529652	1%	70
0.200000002980232	1%	64
0.0299999993294477	1%	59
0.0399999991059303	1%	41
0.0500000007450581	1%	37
0.400000005960464	1%	37
Others	19%	1327
N distinct		445

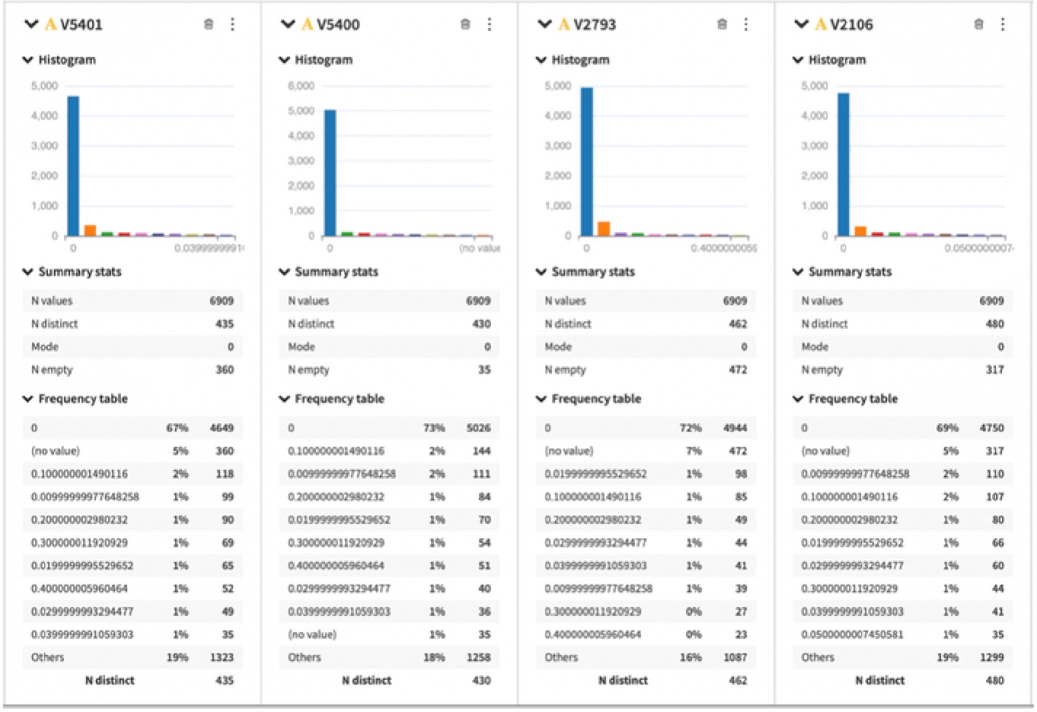
V1801

Summary stats

N values	6909
N distinct	471
Mode	0
N empty	70

Frequency table

0	70%	4845
0.100000001490116	2%	127
0.00999999977648258	2%	126
0.0199999995529652	1%	73
(no value)	1%	70
0.200000002980232	1%	67
0.300000011920929	1%	57
0.0299999993294477	1%	49
0.400000005960464	1%	43
0.0399999991059303	1%	41
Others	20%	1411
N distinct		471



V5401

Summary stats

N values	6909
N distinct	435
Mode	0
N empty	360

Frequency table

0	67%	4649
(no value)	5%	360
0.100000001490116	2%	118
0.00999999977648258	1%	99
0.200000002980232	1%	90
0.300000011920929	1%	69
0.0199999995529652	1%	65
0.400000005960464	1%	52
0.0299999993294477	1%	49
0.0399999991059303	1%	35
Others	19%	1323
N distinct		435

V5400

Summary stats

N values	6909
N distinct	430
Mode	0
N empty	35

Frequency table

0	73%	5026
0.100000001490116	2%	144
0.00999999977648258	2%	111
0.200000002980232	1%	84
0.0199999995529652	1%	70
0.300000011920929	1%	54
0.400000005960464	1%	51
0.0299999993294477	1%	40
0.0399999991059303	1%	36
(no value)	1%	35
Others	18%	1258
N distinct		430

V2793

Summary stats

N values	6909
N distinct	462
Mode	0
N empty	472

Frequency table

0	72%	4944
(no value)	7%	472
0.0199999995529652	1%	98
0.100000001490116	1%	85
0.200000002980232	1%	49
0.0299999993294477	1%	44
0.0399999991059303	1%	41
0.00999999977648258	1%	39
0.300000011920929	0%	27
0.400000005960464	0%	23
Others	16%	1087
N distinct		462

V2106

Summary stats

N values	6909
N distinct	480
Mode	0
N empty	317

Frequency table

0	69%	4750
(no value)	5%	317
0.00999999977648258	2%	110
0.100000001490116	2%	107
0.200000002980232	1%	80
0.0199999995529652	1%	66
0.0299999993294477	1%	60
0.300000011920929	1%	44
0.0399999991059303	1%	41
0.0500000007450581	1%	35
Others	19%	1299
N distinct		480

1.EDA & DATA CLEANING

Conclusions:

- 95% of the values are between 0 and 0.4
- Distinct Values are NA. That is why the column data type is categorical instead of numerical
- In all the cases, the mode is 0. Around 70% of the values are 0 value.

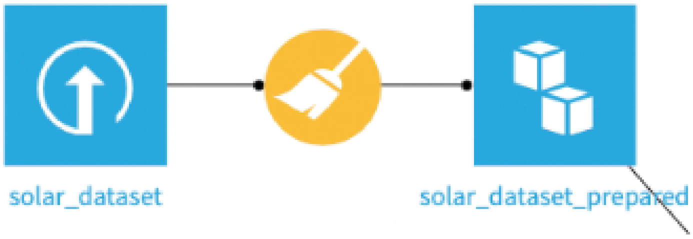
2. DATA CLEANING / PRE-PROCESSING

- **solar_dataset.csv**

Since we will later have to split our dataset for the train / test-val on the basis of our date, we will have to format it accordingly and in a standardised way.
No cleaning is necessary, as we will treat the missing values for the predictor.

Steps

1. Select solar_dataset.csv
2. Visual recipes: Prepare
3. Add a new step
4. Parse to a standard date format
5. Select column: DATE
6. Find with smart date: yyyy/MM/dd
7. Run



1.EDA & DATA CLEANING

2. DATA CLEANING / PRE-PROCESSING

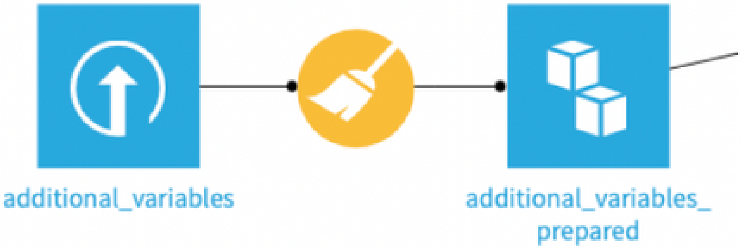
- additional_variables.csv

As we see previously, in this dataset we will need to:

- Adjust the variable types
- Treat missing data
- Identification of atypical data (outliers)

Steps:

1. Select solar_dataset.csv
2. Visual recipes: Prepare
3. Add a new step
4. Parse to a standard date format
5. Select column: DATE
6. Find with smart date: yyyy/MM/dd
7. Add a new step
8. Clear cells in all columns if value is not a decimal
    a. Clearing all NAs
9. Add a new step
10. Fill empty cells of all columns with 0
    a. Changing all NAs by 0 (mode)
11. Change variable type:
    a. string/decimal -> double/decimal



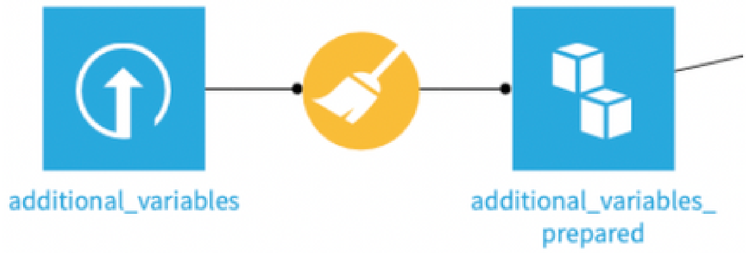
1.EDA & DATA CLEANING

Solar Dataset

Insights:
This dataset has 456 columns and 6909 rows. The first 99 columns indicate the real values of solar production of each of the stations per day. The remaining variables are PCAs (Principal Components Analysis) over the original dataset, given by weather predictors. From row 5113 (01/01/2008), there are NA or missing values, so we focus on predicting the ACME, GOOD, and WYNO stations after 2008.

Preprocessing/Data Cleaning:

- Import the solar_dataset CSV file and infer data types from the schema tab. The station columns are “bigint” type, and the PCA’s are “double” type
- Parse the date column to standard format
- We joined the prepared dataset of solar information with the prepared dataset of additional variables (which also has the date column parsed)
- We joined these two datasets on the date columns using a left join
- From solar dataset, we selected the stations of interest- ACME, GOOD, WYNO and the PCA componentsjjj
- For the additional variables dataset, we included all of the descriptor variables
- We then split this joined dataset into a predictor set which contains the dates from 2008 onwards and a train_val_test, which contains the rows from the start (1994) to the end of 2007 (reference figure 1)
- Perform “analyse” on the train_val_test set



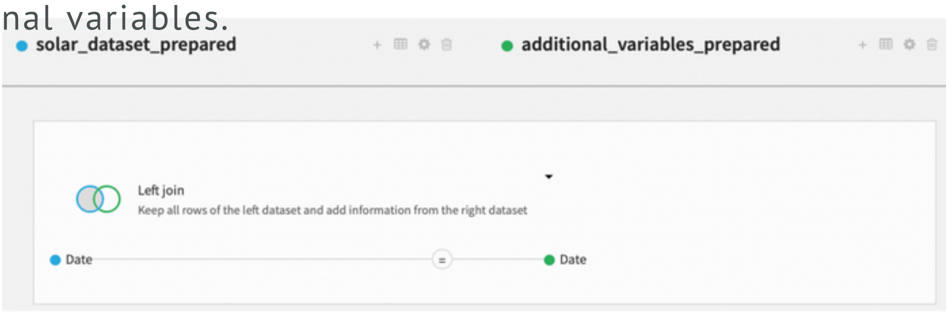
1.EDA & DATA CLEANING

FINAL DATASET

Since we want to have a single dataset to work with, which is all unified (solar_dataset + additional_dataset), we will have to make a join. Including all the possible parameters/variables, that can optimize our models

Steps:

1. Select both datasets (solar_dataset_prepared and additional_variables_prepared)
2. Visual recipes: Join
3. Selected columns:
   a. From solar_dataset we will just select the columns DATE, **ACME**, **GOOD**, and **WYNO** (Our target variables) + all the Principal Components PCs. (PC1, PC2, PC3, ...)
   b. From additional_variables we will select everything except DATE (already selected in solar_dataset)
4. Join: This will select the date with our three variables to predict, plus all additional variables.



● solar_dataset_prepared

● additional_variables_prepared

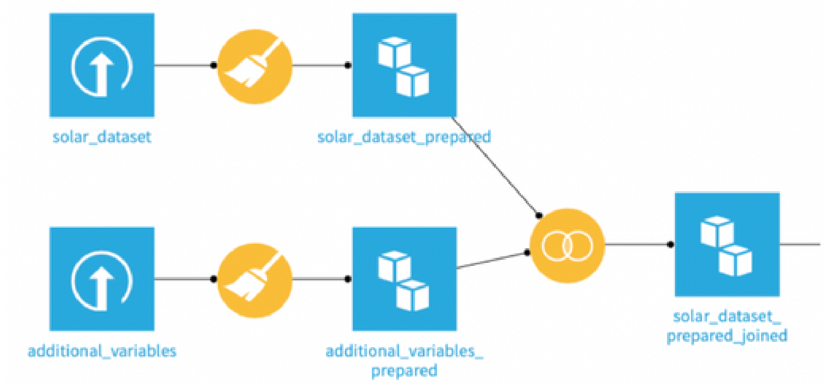
Left join

Keep all rows of the left dataset and add information from the right dataset

● Date = ● Date

5. Run

So the final Dataiku WorkFlow is...



2. MODELLING

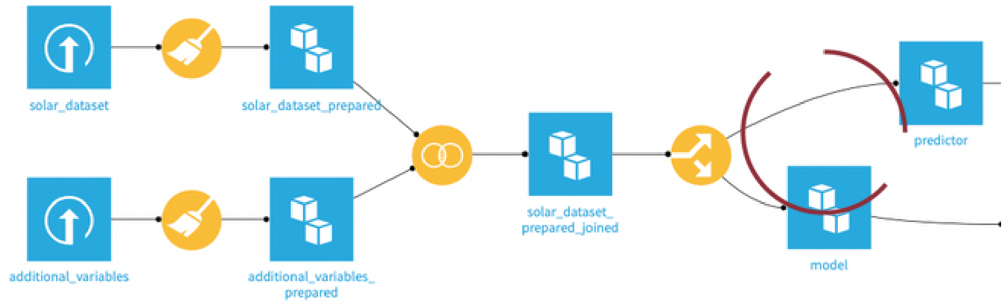
According to the first part of the assignment, once all the data cleaning, EDA & pre-processing phase is done, we will work with our final dataset for the predictions of our 3 variables: ACME, GOOD and WYNO.

It is important to note that our final dataset, as of 01/01/2008, does not have any value for any of the stations. That said, to train and validate our regression model, we will have to make a split according to the date:

- 1st split : **model** - all data from 01/01/1994 to 31/12/2007. On this dataset, we will train the model.
- 2nd split: **predictor** - all data from 01/01/2008. This dataset is the one with all the missing values, and where we will apply our winning model (trained on the dataset:model) to predict our values from 01/01/2008 to 30/11/2012.

<u>**STEPS**</u>

1. Select solar_dataset_prepared_joinded
2. Visual Recipes: Split
    a. Output: Add model and predictor
    b. Create Recipe
3. Define filters
    a. location: model
    b. keep only rows that: following conditions
    c. WHERE: Date is before 01/01/2008
    d. all other values: predictor



2. MODELLING

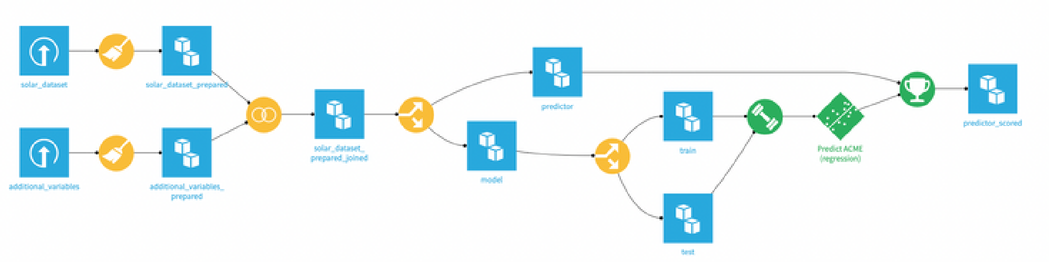
Once we have divided our final dataset. We will focus on the dataset: **model**, to train and validate our regression models.

- Train: Part of our dataset, to extract patterns and predict the target variable.
- Validation: Used to select the best trained model, performing the parameter adjustment or metamodeling.
- Test: Provides the best model + the actual error expected

We will do a split: train/test-val (80-20)%. This means that exactly 80% of data from 01/01/1994 to 31/12/2007 will be in the train dataset (approx. 11 years).

STEPS

1. Select model dataset
2. Visual Recipes: split
    a. Output: train and test
    b. Create Recipe
3. Dispatch percentiles of sorted data on output datasets
    a. Sort according to: Date
    b. 80% train / 20% test
4. Run



From this point, we will change some parameters to improve the models of each of the variables to be predicted, but always working from these two datasets to train and validate (maximum optimisation).

Once our winning model is selected, we will apply it to the dataset: predictor (with blank values), for the final prediction.

2. MODELLING

ACME MODEL

2. MODELLING

GOOD MODEL

TARGET + PCS

- LASSO - 2.52

STEPS

1.

2. MODELLING

WYNO MODEL

Text text text
Hello hello hello