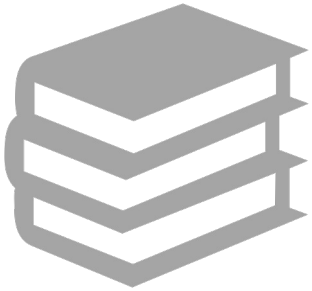
# PREDICTIVE ANALYTICS And DATA MINING

PADM Spring 22 MASTERS EPITA


- Axel ULLERN

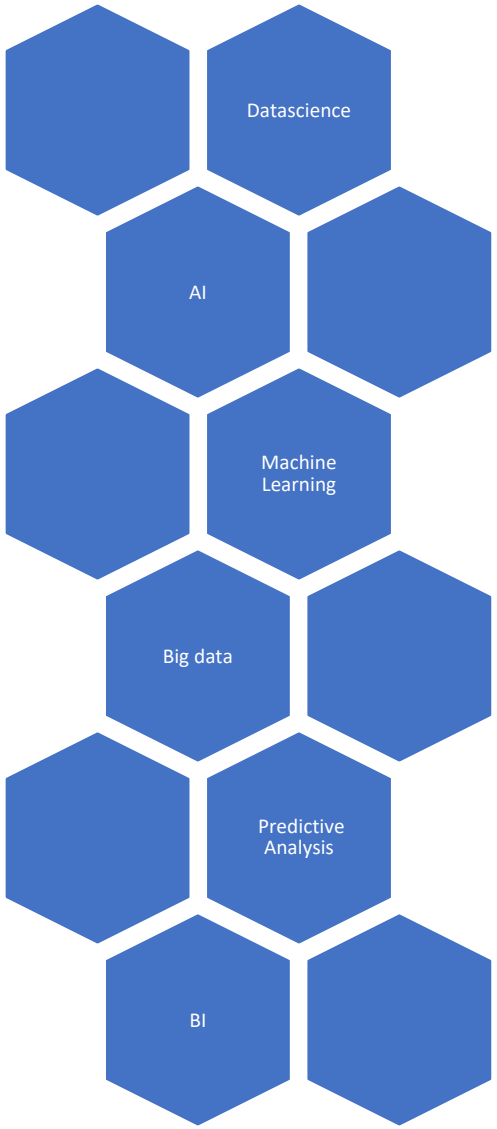
# INTRODUCTION

FUNDAMENTALS

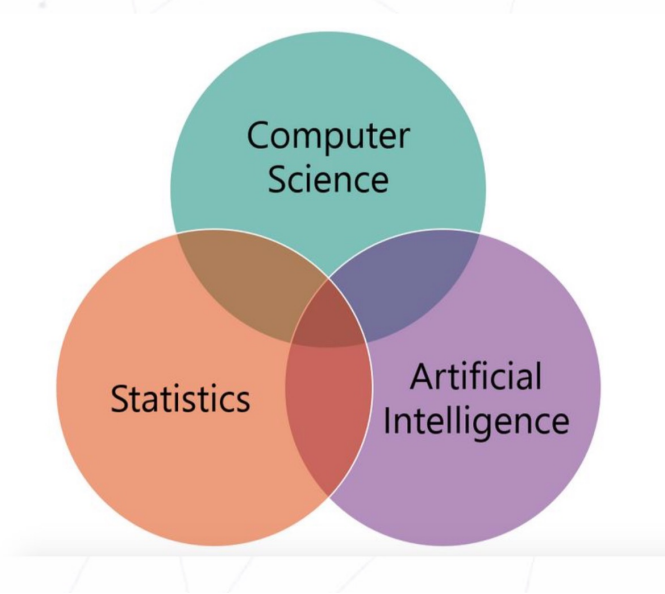


VOCABULARY
USE CASES OF DATASCIENCE

# PIECES OF VOCABULARY



# And some of the definitions




**Datascience** = Create Business Value through Analysis of datas Combining those 3 Technologies

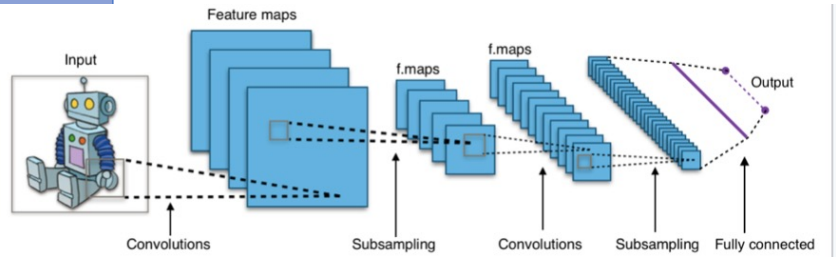
**Machine Learning** = Machine Learning is an Artificial intelligence technology allowing computers to learn without having been explicitly programmed for this purpose. ...

In fact, **Big Data** is the essence of Machine Learning and it is the technology that fully exploits the potential of Big Data.

**Deep Learning** = Class of machine Learning using sophisticated algorithms To solve complex problems

# Deep Learning

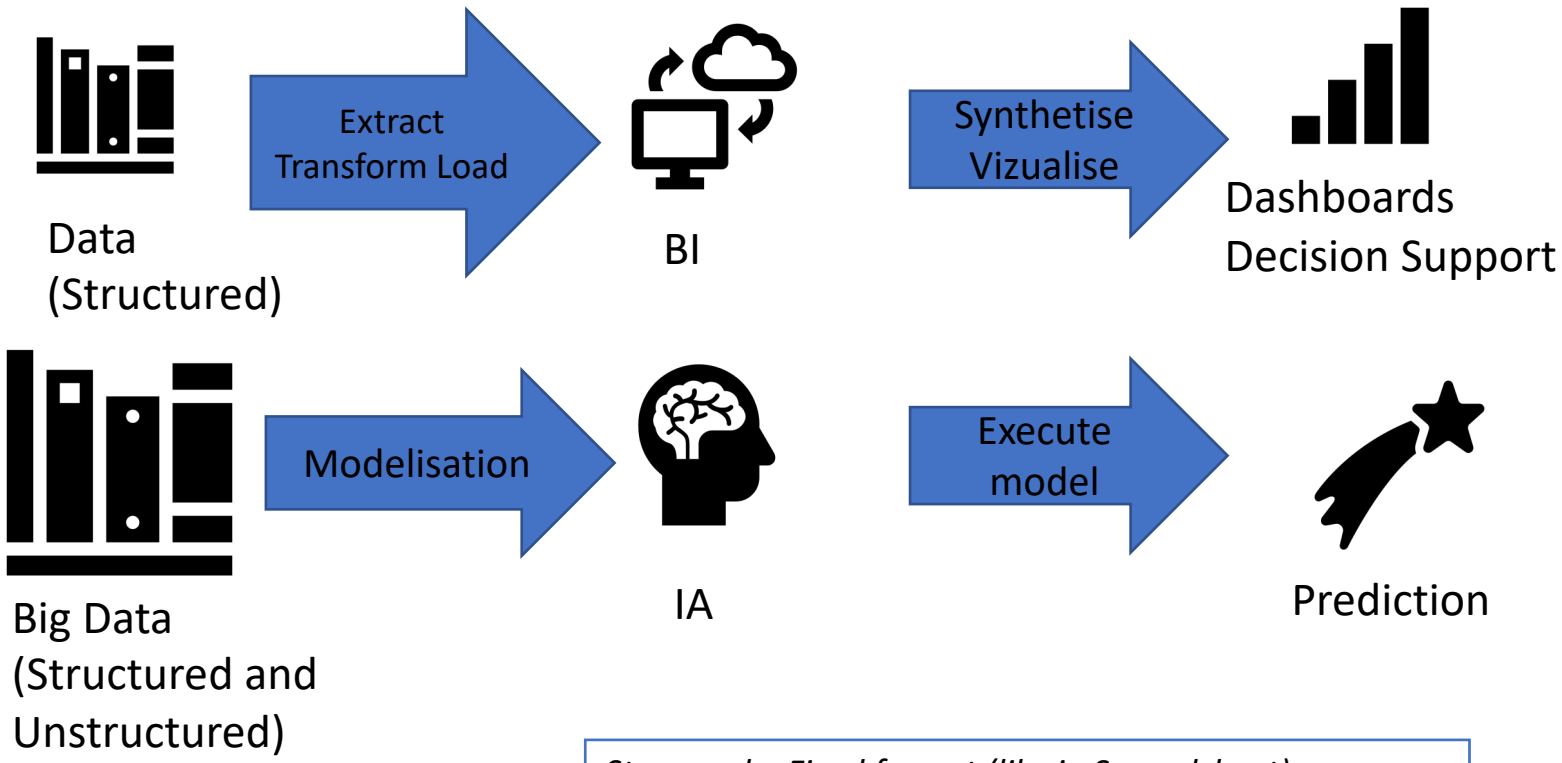
Deep learning uses hidden layers of artificial neural networks, and series of complex propositional calculations. Deep learning algorithms oppose shallow learning algorithms because of the number of transformations performed on the data between the input layer and the output layer.




**Many applications :**

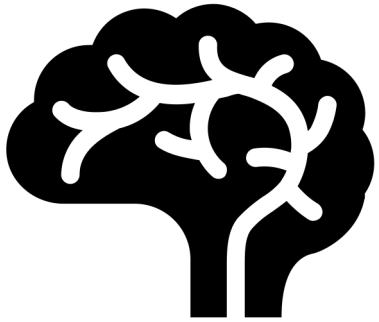
Speech recognition
Natural Language Processing
Image Analysis
Illnesses Diagnosis,
……

# BUSINESS INTELLIGENCE VS AI

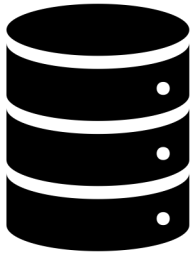



*Strucured = Fixed format (like in Spreadsheet)*
*Unstructured = Diverse format (Voice, image, text)*

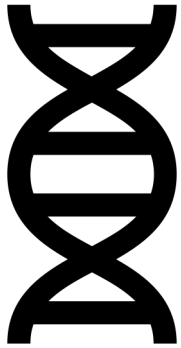
# WHY SO MANY DATAS ?



- Wallmart processes more than one million transactions per hour in its stores
- In medicine, epidemiology and genetics we operate billions of data
- Web logs easily get billions of data
- Facebook stores pictures by billion
- The IOT generate huge amounts of data



- Finally .. To make reliable predictions requires a lot of data without datas
- **IA or Machine Learning would not exist**

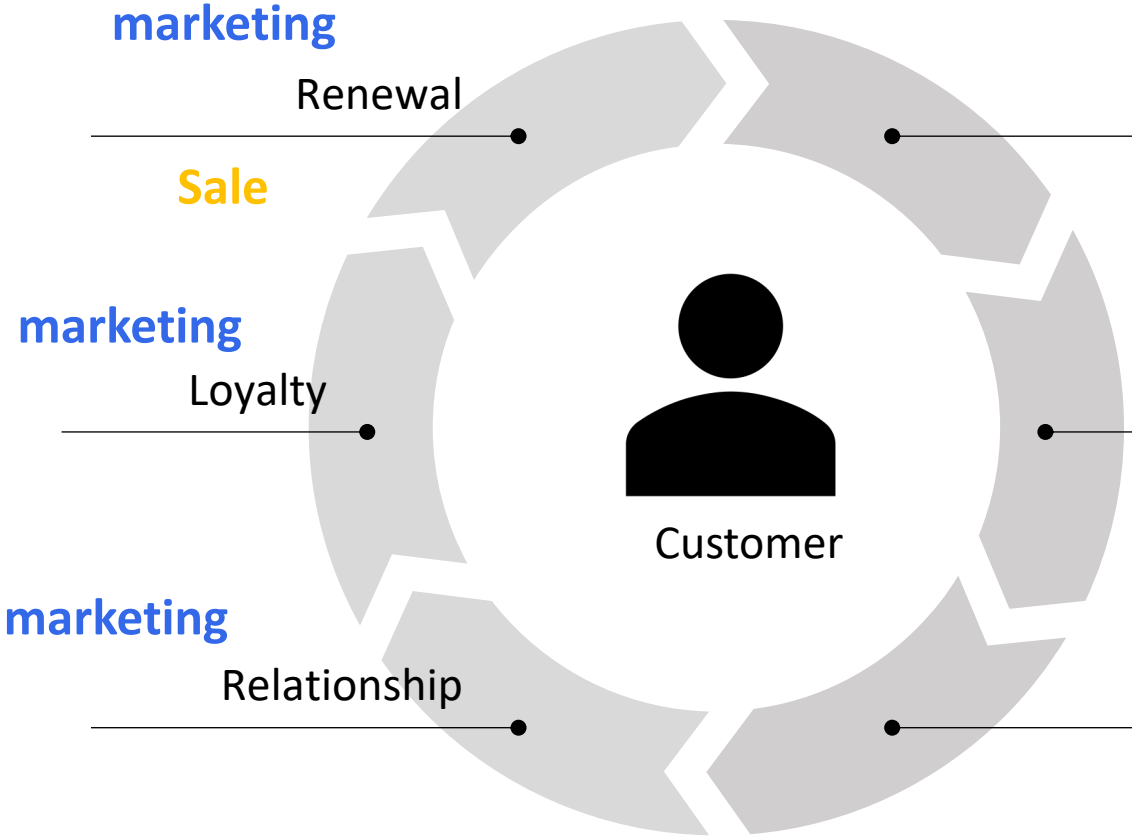


# CUSTOMER EXPERIENCE: DATA SOURCE

Awareness: advertising

Website (Pull) →contacts /Leads

marketing



# WHICH PREDICTIONS?

Health : illnesses, Tumors,…

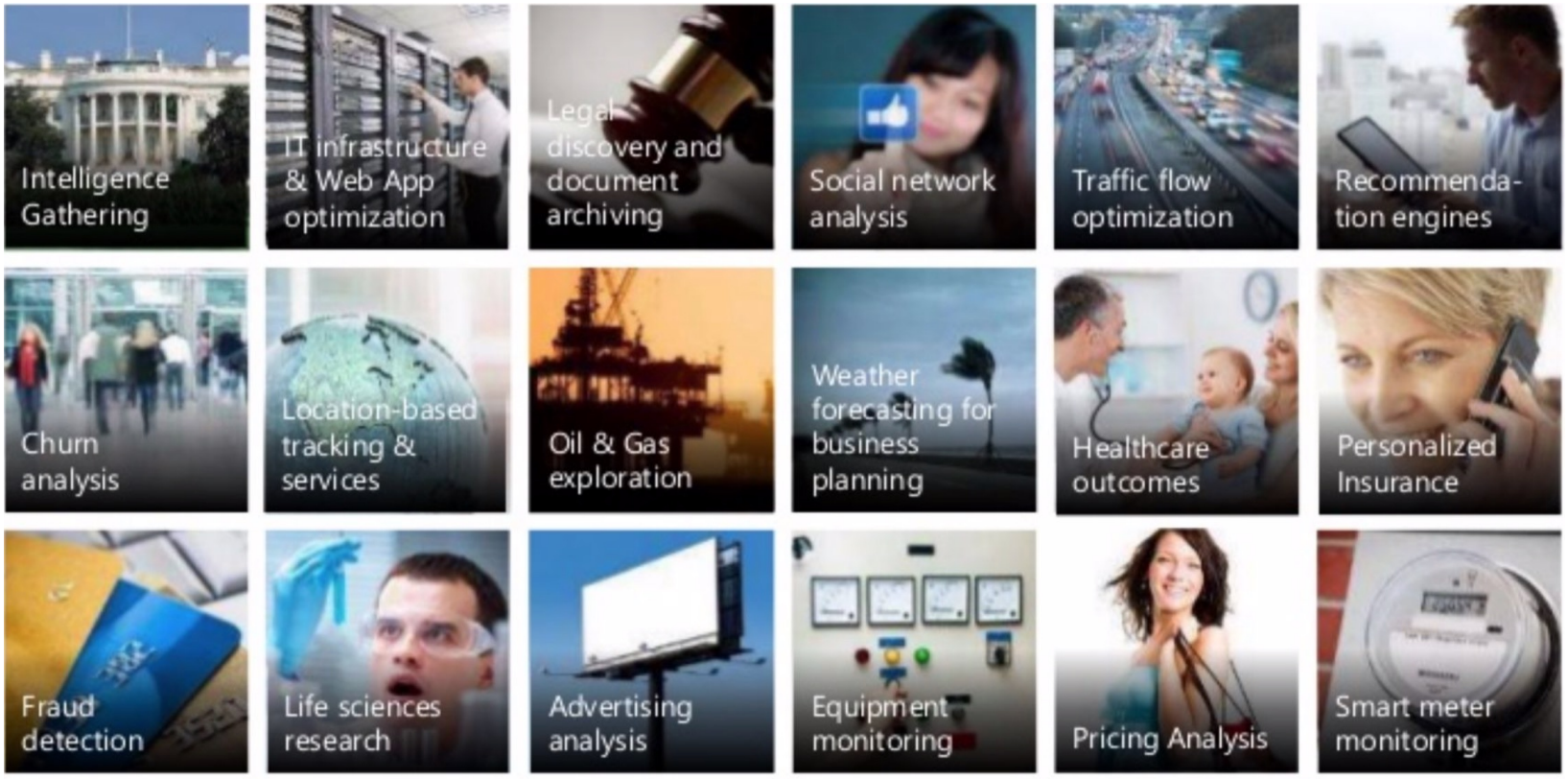
Customers : Churn, recomandation engines,advertising..

Industry 4.0 : Anticipate failures, Predictive maintenance

Banks : Risk management

Security : Fraud detection , intrusions

# MAIN USES CASES for DATASCIENCE



# DATA IN REAL LIFE

HOW COMPANIES PROCESS DATA ?

WHAT IS A DATA ANALYSIS PROJECT?

# FIRST COMPANIES HAVE TO HAVE A BUSINESS PROBLEM TO SOLVE OR BUSINESSES ANSWERS TO FIND

- Why  customers quit their contract ?
- Should we grant a loan to this customer ?
- Will our customers buy this new product ?
- When this machine will fail ?
- Why employees move to competitors ?

*Big Data and AI will probably provide answers*
*But will never take decision : companies have to take decisions*
*Based on those answers*

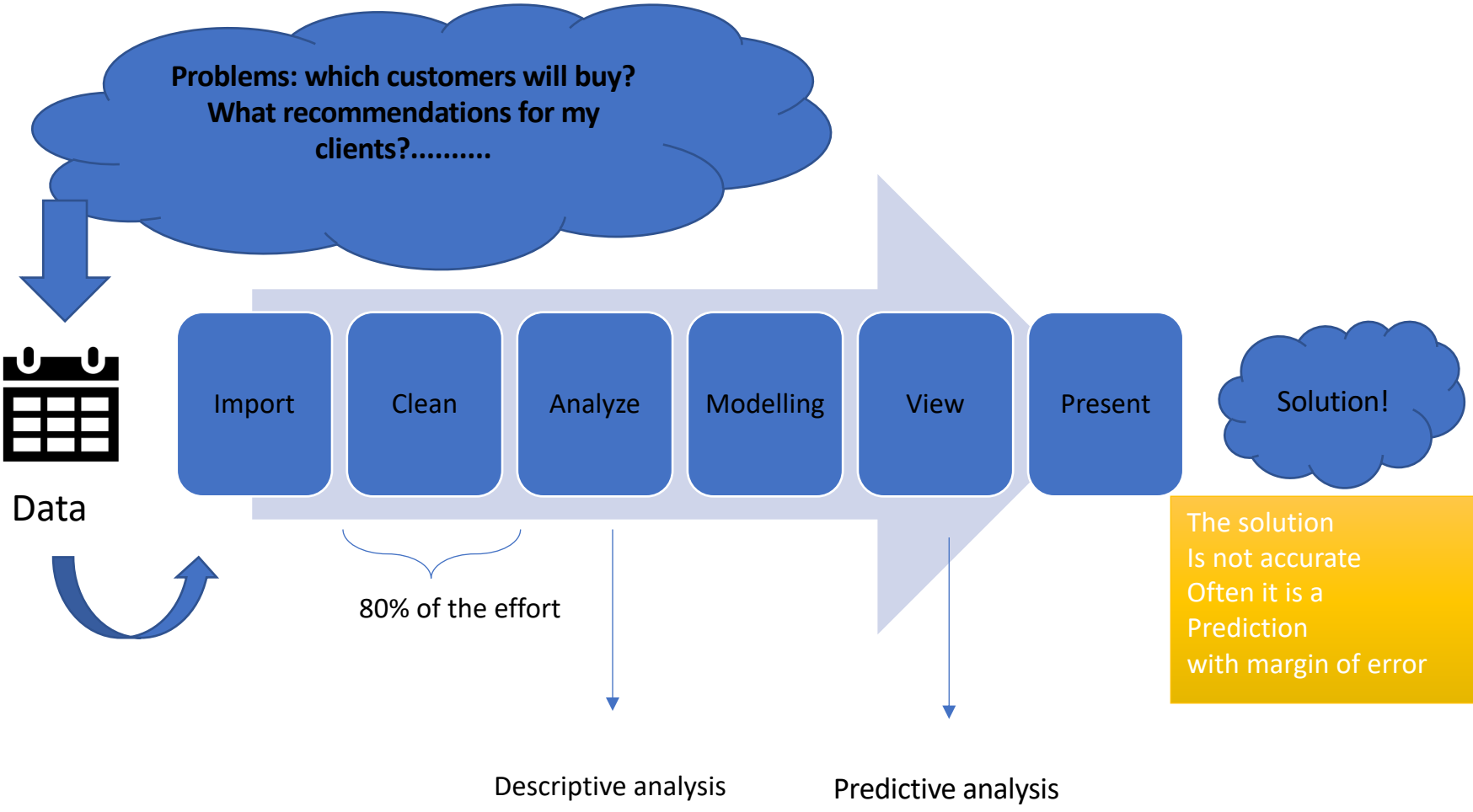
# Then they have to collect DATA FROM VARIOUS SOURCES

# THE PREP PHASE TAKES LONG TIME !

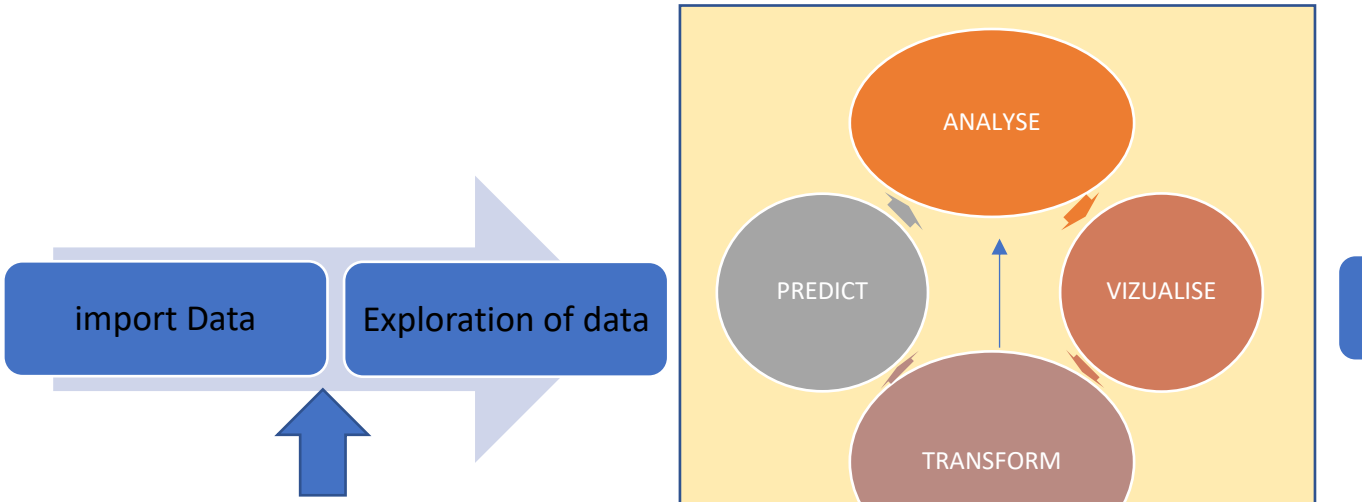
80% of the work to be done: data set preparation

20% analysis

# From the problem to the solution



# THE ANALYTICAL PROCESS OF DATA



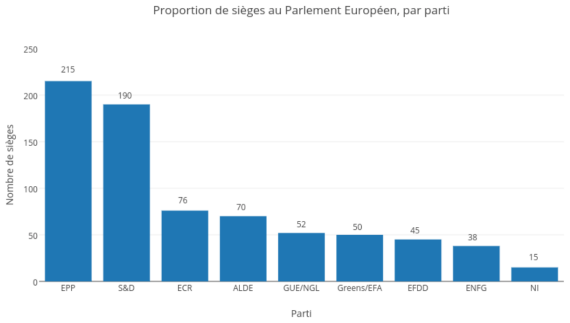

Business Issue to Solve

You can do the four steps above
Or bypass prediction
And do 3 steps : Simple Analysis

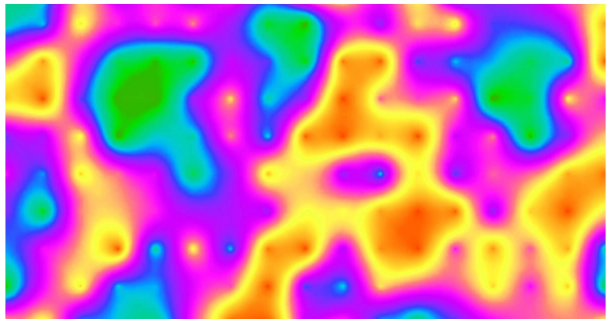
Business Issue Solved !

# FORGET EXCEL !
# THE  VOLUME CHANGE THE SCALE

When a small volume
Data (100-1000) to be visualised : you can have this like in Excel spreadsheet




But when you have a huge Volume ( Millions or Billions) Data you get often this



# IMPORTANCE OF CHARTS AND PLOTS in BIG DATA

" *IT IS QUITE IMPOSSIBLE TO* ***VIZUALISE LARGE AMOUNT OF*** *DATA (BIG DATA) SO WE NEED* ***synthesized the data*** *USING* ***STATISTICS*** *AND* ***GRAPH*** "

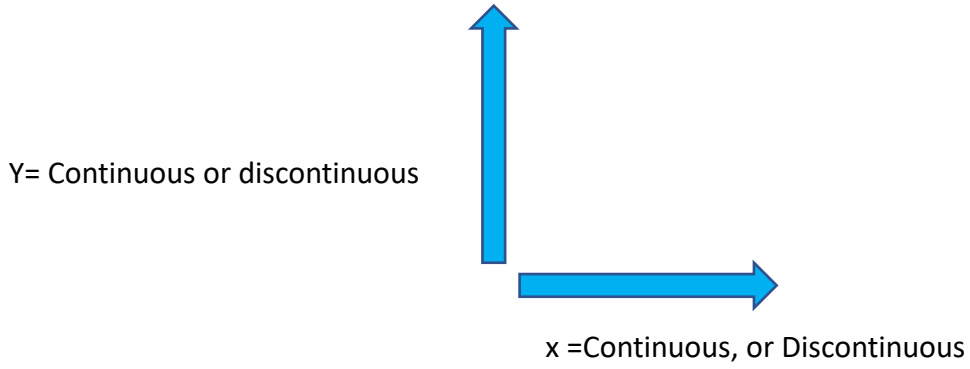
Sometimes the graph is a cloud of points unreadable point cloud on which
It is **impossible to identify a trend so you have to transform and reduce the volume**

# 2 DATA TYPES

1 **Continuous data** They can take any value
(integer or decimal).
Eg temperature, pressure, humidity, sales, etc ..
Or: logs onto a website

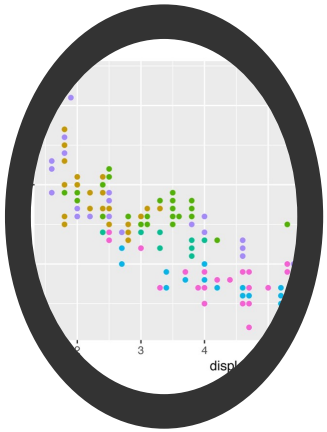
2 **Discontinuous or categorical data** They have predetermined values
or fall into categories :
(BAC + 1 ,MASTERS, ...) or the questionnaire responses (yes, no, do not know)
Or socio-professional categories ....

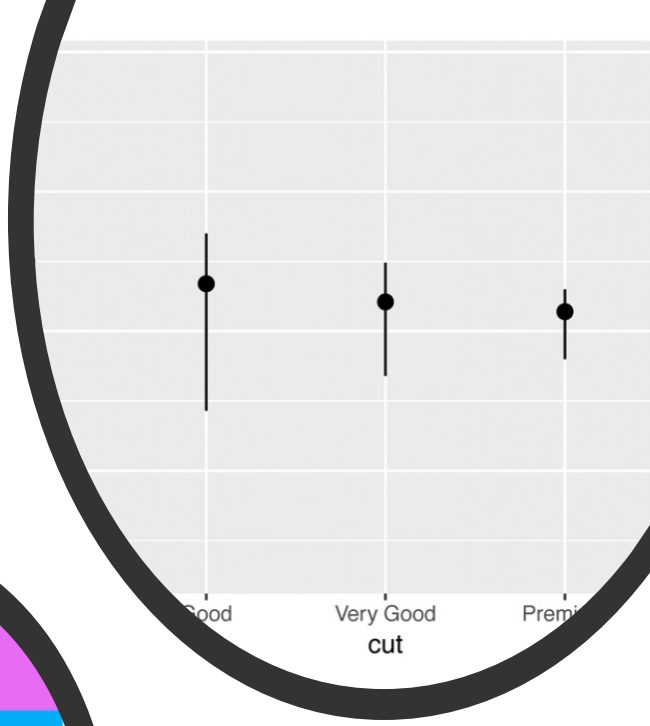
***Consequence*** *: We get different kinds of plots to represent series of data*

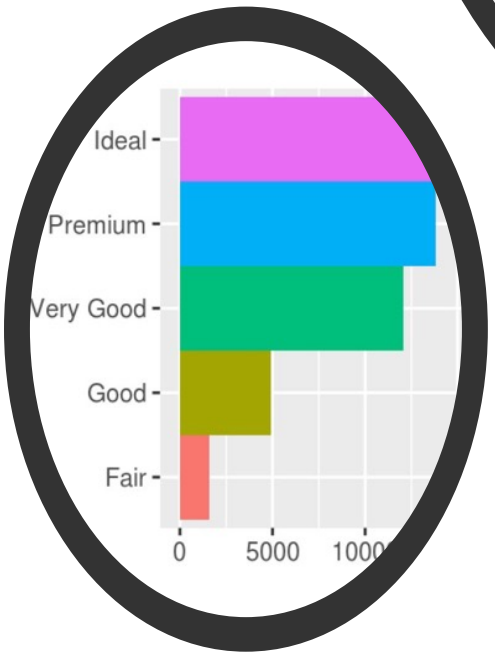


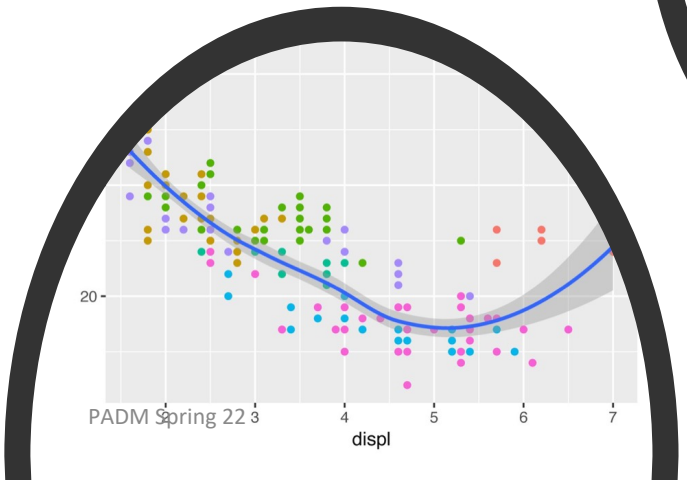
# EXEMPLES

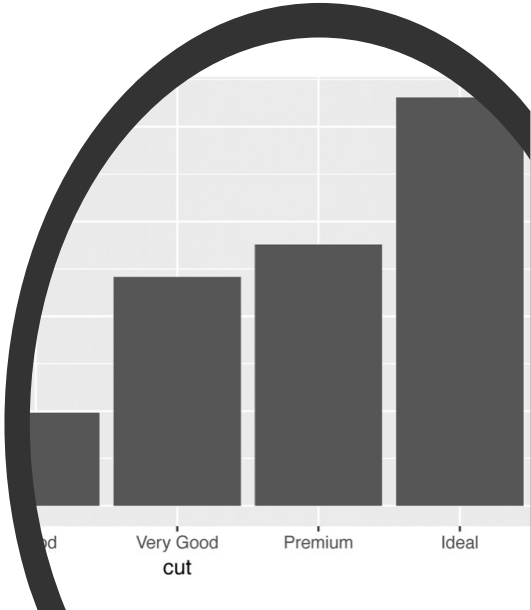
- In all these plots (about diamonds)which kind of
- Data are represented ?
- How much series of data can be plotted
- On the same graph ?



# GENTLE REMINDERS OF STATISTICS

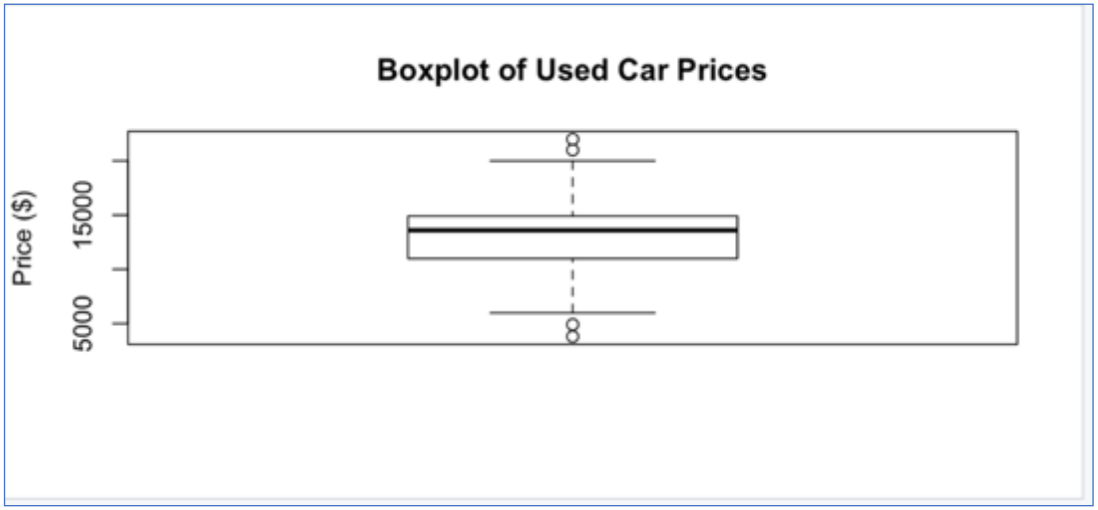
MEDIAN vs MEAN

BOXPLOT

CORRELATIONS

# STATISTICS TOOLS

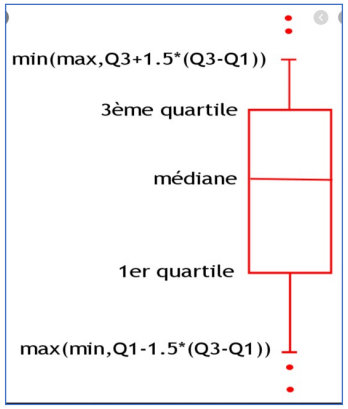
- Understanding the data
- Statistical tool:
- Means,
- Medians,
- Variance and Covariance , etc.
- Highlight correlations between variables



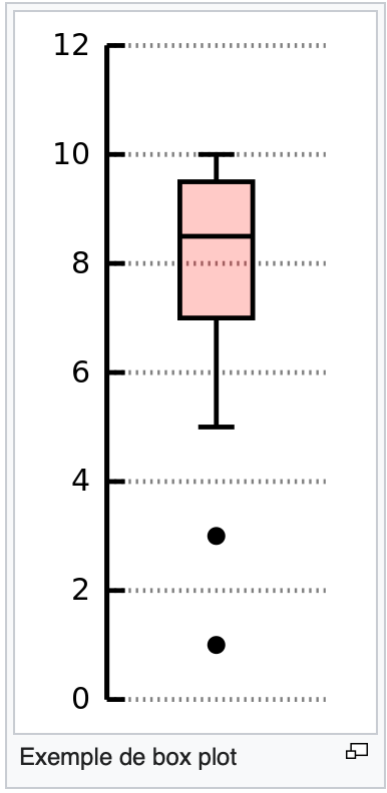
# Median and Boxplot

The series of numbers below must first be sorted in ascending order
Then the boxplot is used to represent the statistical distribution of the series

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ (non triés) | 9 | 6 | 7 | 7 | 3 | 9 | 10 | 1 | 8 | 7 | 9 | 9 | 8 | 10 | 5 | 10 | 10 | 9 | 10 | 8 |
| $x_{(i)}$ (triés) | 1 | 3 | 5 | 6 | 7 | 7 | 7 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 10 | 10 | 10 | 10 | 10 |




Median = middle of the series here average between
The 10th and 11th so **8.5**



# WHAT THE BOXPLOT SHOWS ?

When the box plot is **short** it implies that much of your data points are similar, since there are many values in a small range

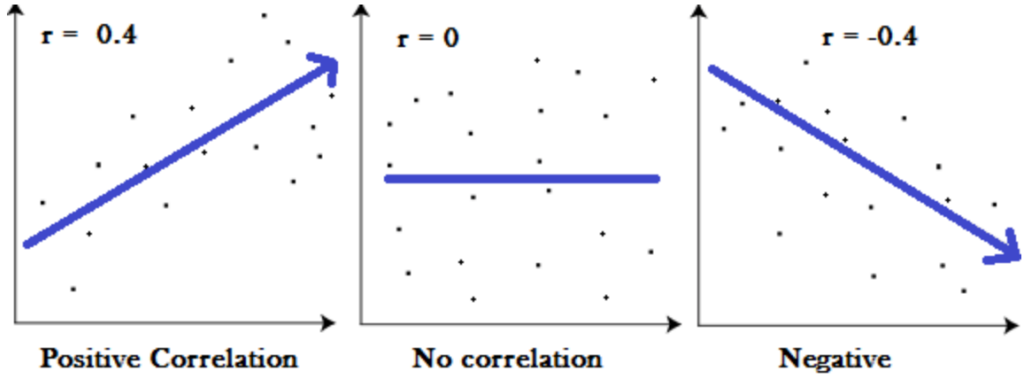
When the box plot is **tall** it implies that much of your data points are quite different, since the values are spread over a wide range

If the median value is closer to the **bottom** then we know that most of the data has lower values. If the median value is closer to the **top** then we know that most of the data has higher values. Basically, if the median line is not in the middle of the box then it is an indication of **skewed**data.

Are the whiskers **very long**? That means your data has a high **standard deviation** and **variance** i.e the values are spread out and highly varying. If you have long whiskers on one side of the box but not the other, then your data may be highly varying only in one direction.

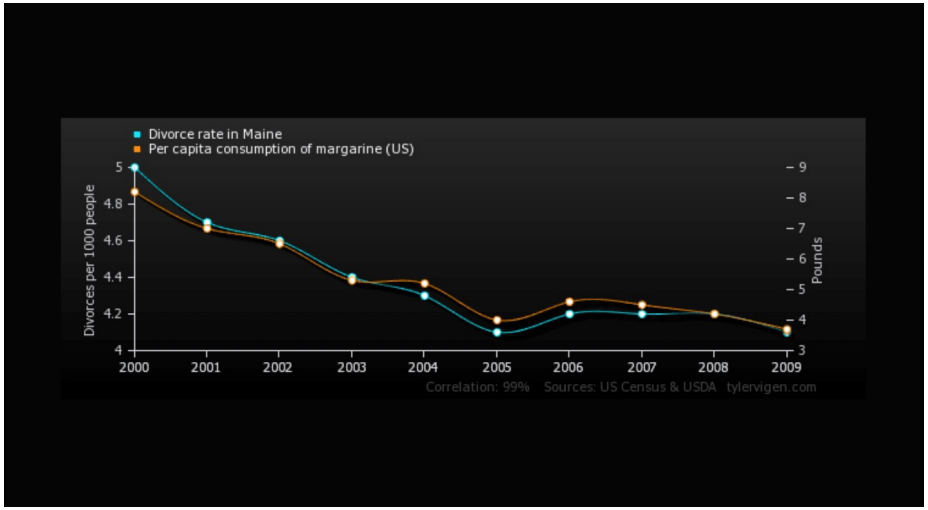
# CORRELATION

- What is correlation? Correlation is **a statistical measure that expresses the extent to which two variables are linearly related** (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.



# CORRELATION AND CAUSATION

- Correlation or causation?
- A common reasoning error is to say, "X and Y are correlated, so X causes Y". We then confuse correlation and causation
- It could also be that:
  - Y causes X
  - X and Y have a common cause Z
  - X and Y are accidentally linked but have no link of causality.




Based on this graph
One could pretend that
Margarine consumption
Cause divorces !

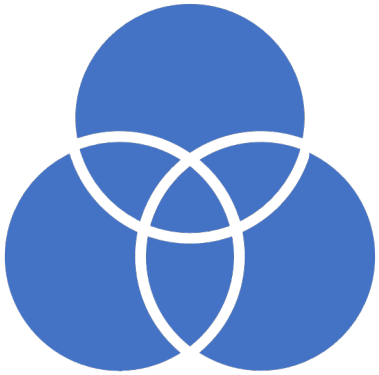
# If you have 2 continuuous variables : check the correlation

If we find a strong correlation >0.6 or <-0.6 then we ask ourselves the question: causality or not?
It is then necessary to dig to check if there is a cause or if it is chance.

Some correlations are obvious and have no interest e.g. ice cream sales
And outside temperature
Others may have a real interest:
Walmart found a correlation between sales of baby diapers and sales of Beer packs!

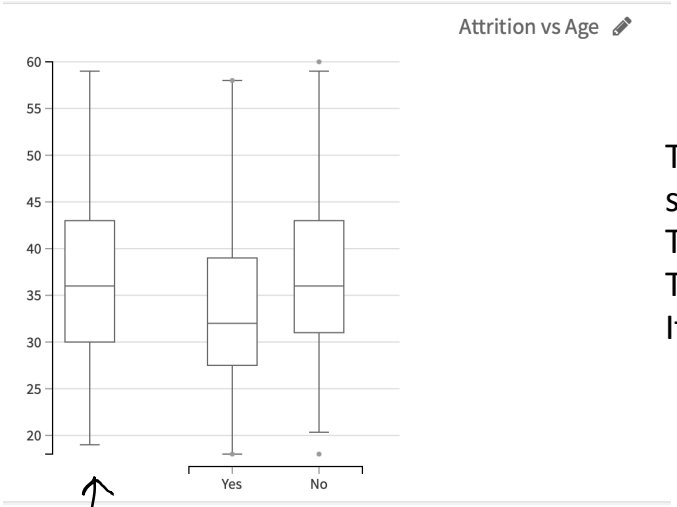

# Is there a correlation between a continuous and categorical variable?

- NO, IT MAKES NO SENSE! BOTH VARIABLES MUST BE CONTINUOUS TO CALCULATE THE CORRELATION
- NEVERTHELESS, IT IS INTERESTING TO MEASURE THE RECIPROCAL "INFLUENCE" OF A CONTINUOUS VARIABLE AND A CATEGORICAL VARIABLE.
- BUSINESS (HR) QUESTION: IS LEAVING OR STAYING IN THE COMPANY RELATED TO AGE?
- LEAVE THE COMPANY = CATEGORICAL VARIABLE (YES OR NO)
- AGE = CONTINUOUS VARIABLE
- HOW TO DO IT?
- ANSWER: DRAW TWO BOXPLOTS ONE CORRESPONDING TO THE EMPLOYEES WHO HAVE LEFT THE COMPANY THE OTHER TO EMPLOYEES WHO LEFT



# Attrition and Age: what relationship?

*Attrition = Leaving the company  Attrition = Yes employee left , No= employee stayed*




All employees

Employees who stayed

Employees who have left

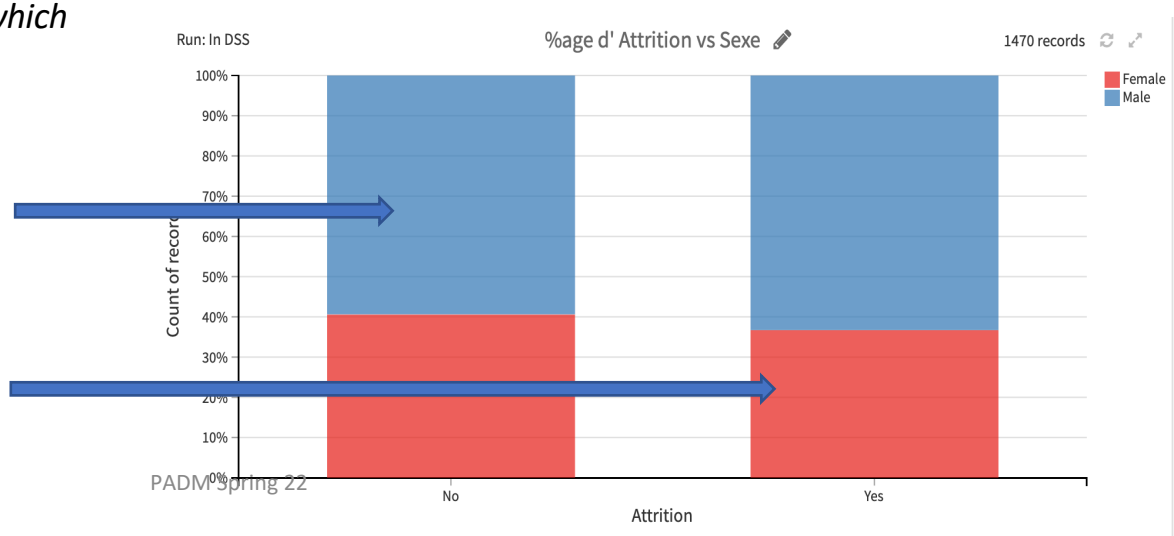
The positioning of the two boxes
shows that employees who have left
The company is overall younger
That those who stayed
It is then necessary to look for the causes of this situation

# Links between two categorical variables

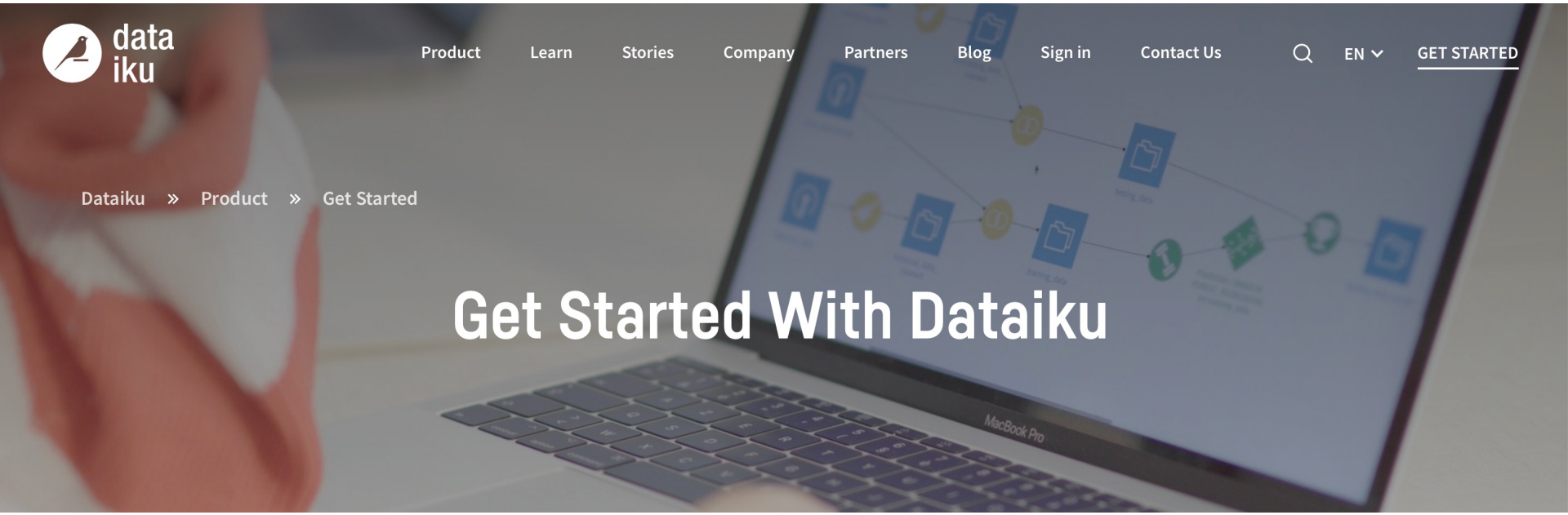
- **Business Question: Are women or men most likely to leave**
- **The company?**
- **Male or Female = Categorical variable**
- **Exit or stay = Categorical variable**
- **How to do it? : a graph with two bars one for employees who left the other for those**

**Who remained, inside each bar two colors for H or F, height of the bars = number of employees**
**We compare the colored areas between the two bars**

*By comparing the 2 blue or red zones*
*From each bar we see that there is a little less*
*Of Women than Men in proportion to which*
*Leave the company*

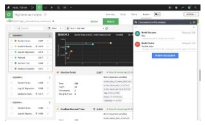
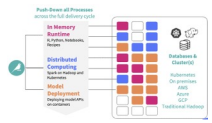


# And Now…….


data iku
Product
Learn
Stories
Company
Partners
Blog
Sign in
Contact Us
EN
GET STARTED
Dataiku » Product » Get Started
Get Started With Dataiku
MacBook Pro

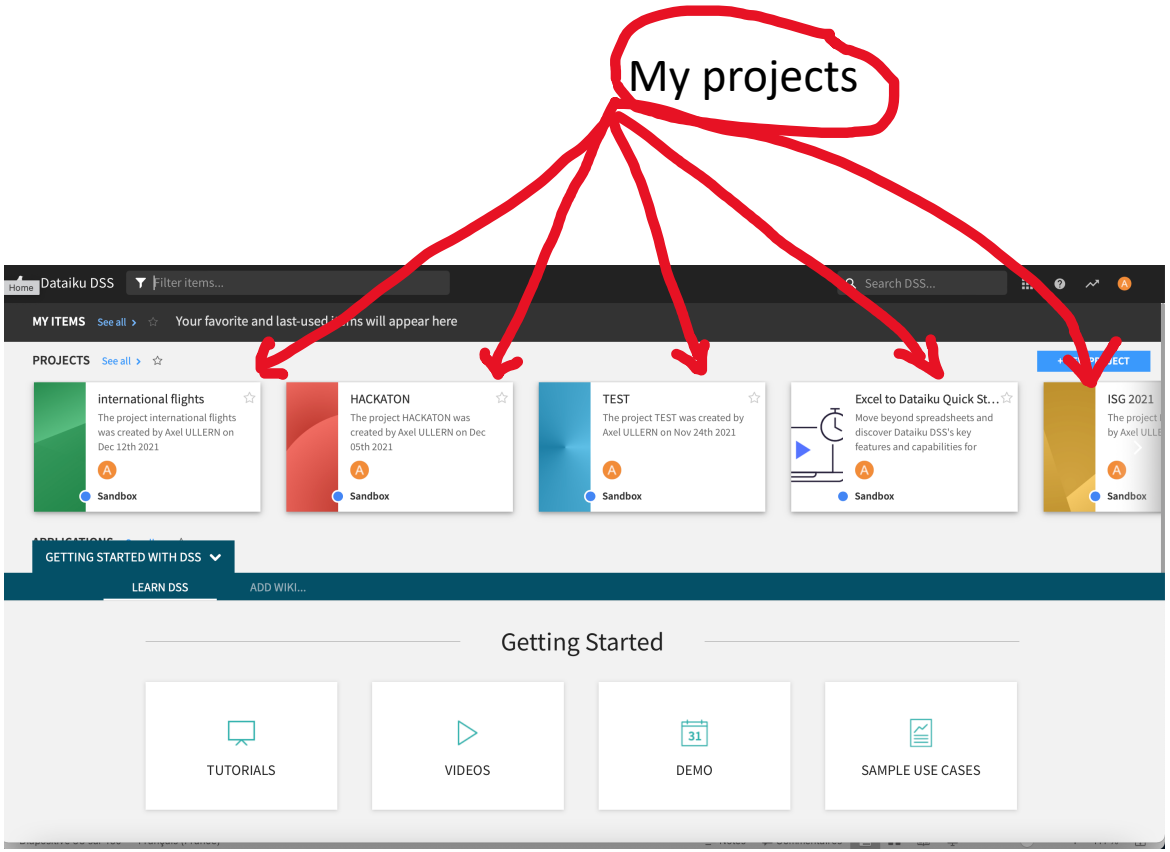
## Key Company Milestones

# A Brief History of Dataiku

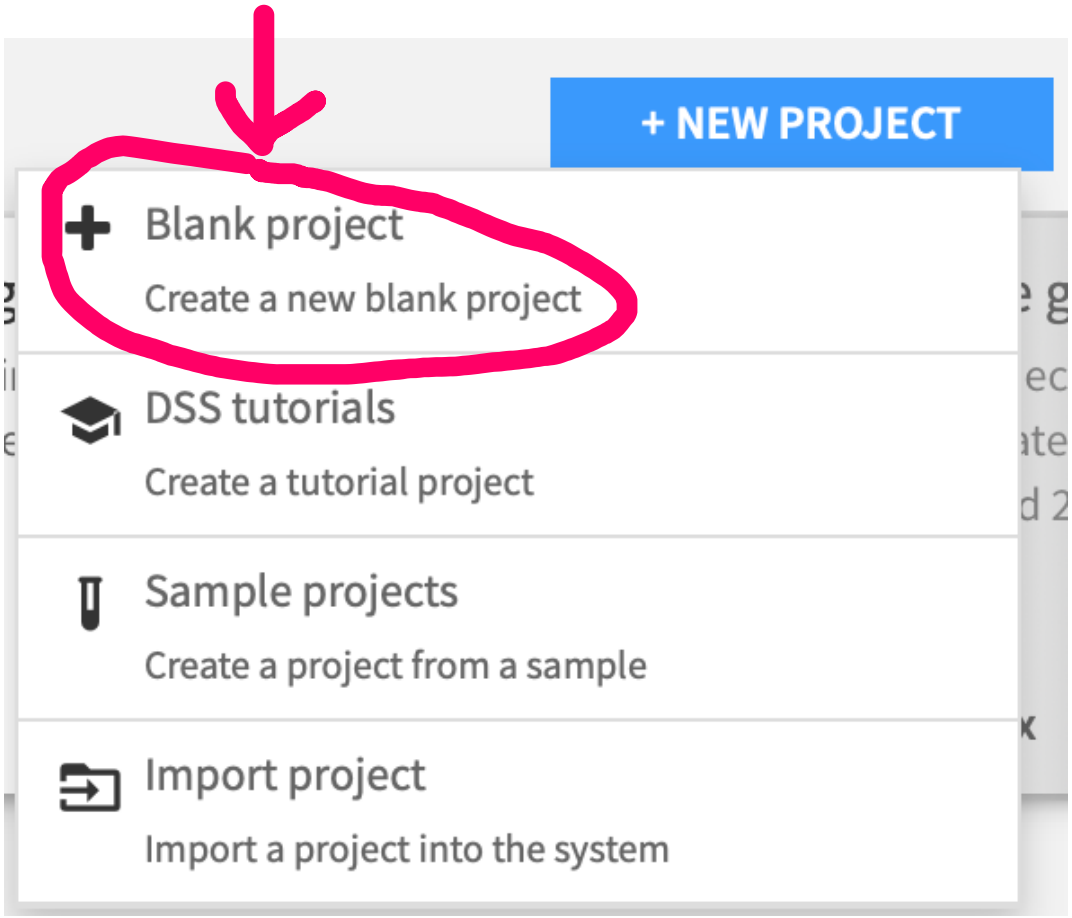
| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|---|---|---|
| COMPANY | **Dataiku Created** | **20th employee** | Office in New York | **$4M raised** | Office in London | **100th employee** | **200th employee** | **400th employee** |
| | | | | **$14M Series A** | | **$28M Series B** | 7 new offices in EMEA, APAC, Americas<br>**$101M Series C** | **$1.4B valuation** |
| CUSTOMERS | First customer | | First 10+ users customers | 50th customer<br>First 100+ users customers | EGG CONFERENCE | 100th customer<br>1st CAB Customer Advisory Board<br>First 500+ users customers | 300th customer<br>First 1000+ users customers | 1000+ EGG Visitors |
| SOFTWARE |  Beta |  version 1<br>1st tool integrating data prep and ML |  version 2<br>Real-time collaboration Spark integration |  version 3<br>Scenarios, automation and version control |  version 4<br>Advanced scalability, cloud support |  version 5<br>Deep learning, Containerization, Project wikis |  version 6<br>Enterprise Scale Elastic AI with Kubernetes |  version 7<br>Explainable AI and Advanced Visual Statistics |

- GETTING STARTED WITH THE PLATFORM
- CREATING A PROJECT
- DATA IMPORT

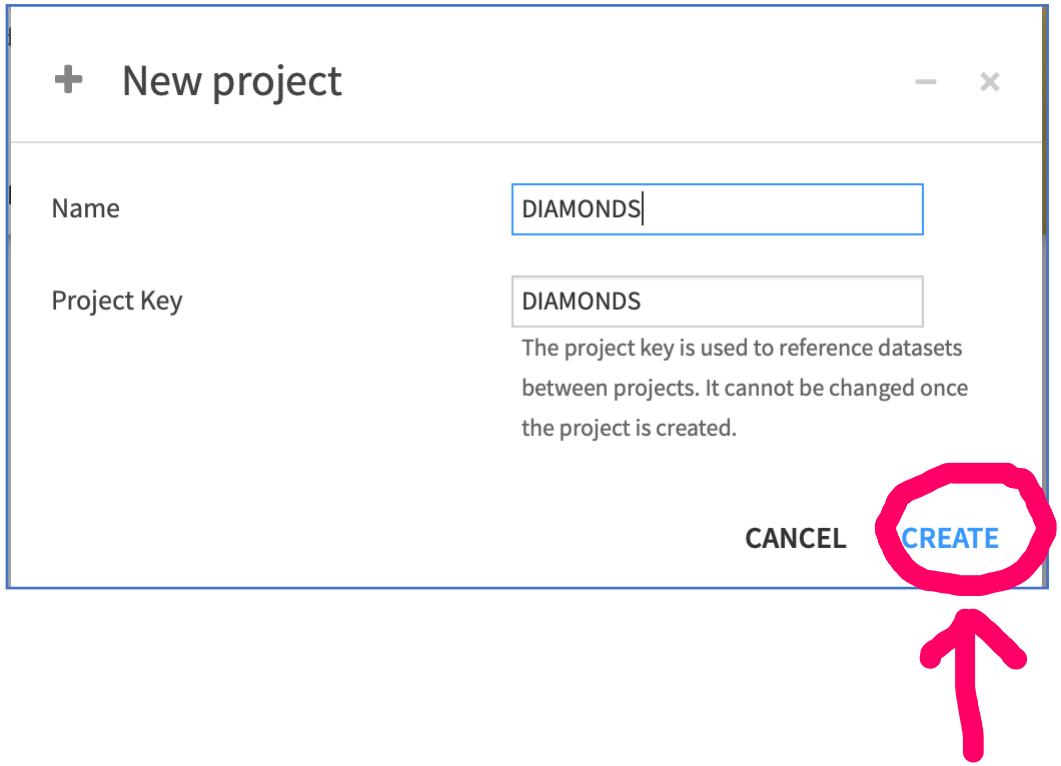
# Home Page

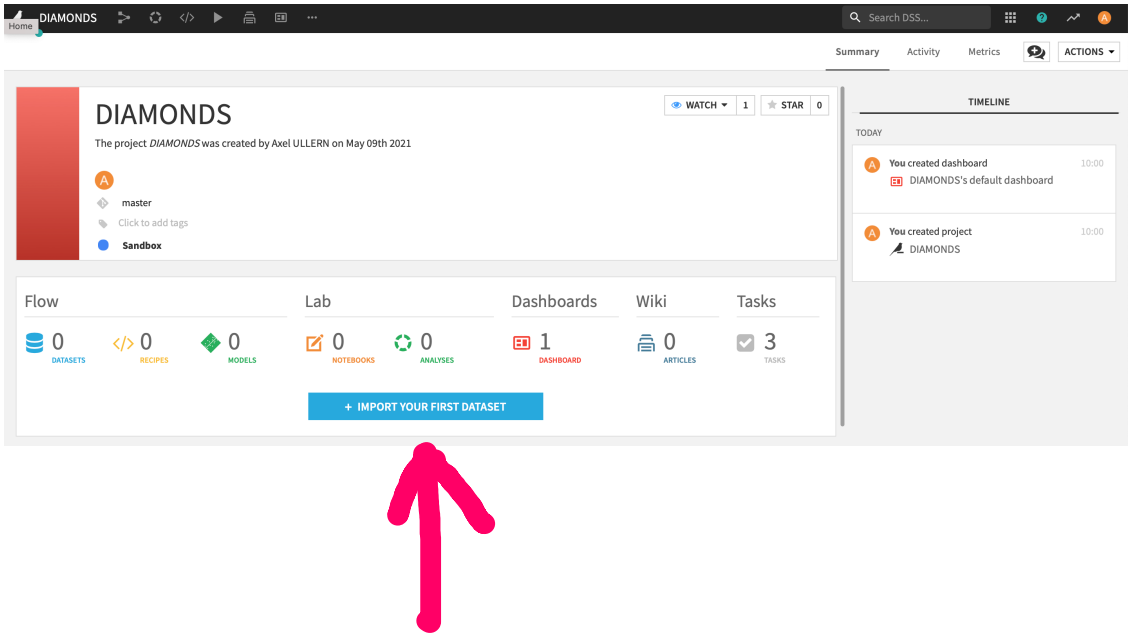

My projects
Dataiku DSS
Filter items...
Search DSS...
MY ITEMS
See all
Your favorite and last-used
ns will appear here
PROJECTS
See all
international flights
The project international flights was created by Axel ULLERN on Dec 12th 2021
Sandbox
HACKATON
The project HACKATON was created by Axel ULLERN on Dec 05th 2021
Sandbox
TEST
The project TEST was created by Axel ULLERN on Nov 24th 2021
Sandbox
Excel to Dataiku Quick St...
Move beyond spreadsheets and discover Dataiku DSS's key features and capabilities for
Sandbox
ISG 2021
Sandbox
GETTING STARTED WITH DSS
LEARN DSS
ADD WIKI...
Getting Started
TUTORIALS
VIDEOS
DEMO
SAMPLE USE CASES

# CREATING A PROJECT



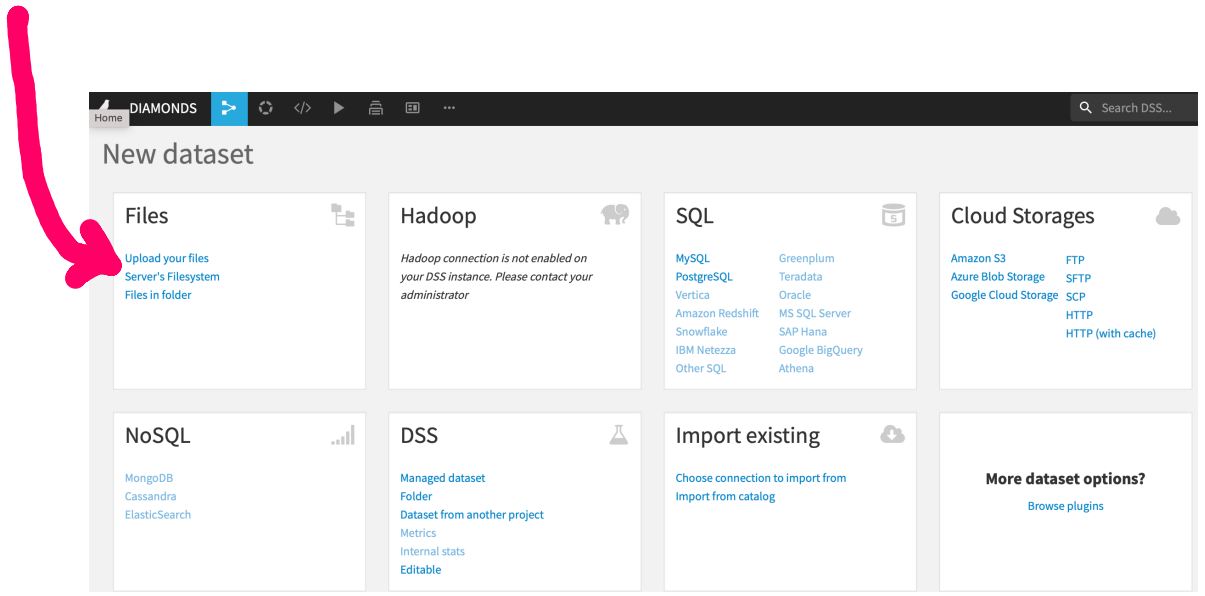
# NEW PROJECT

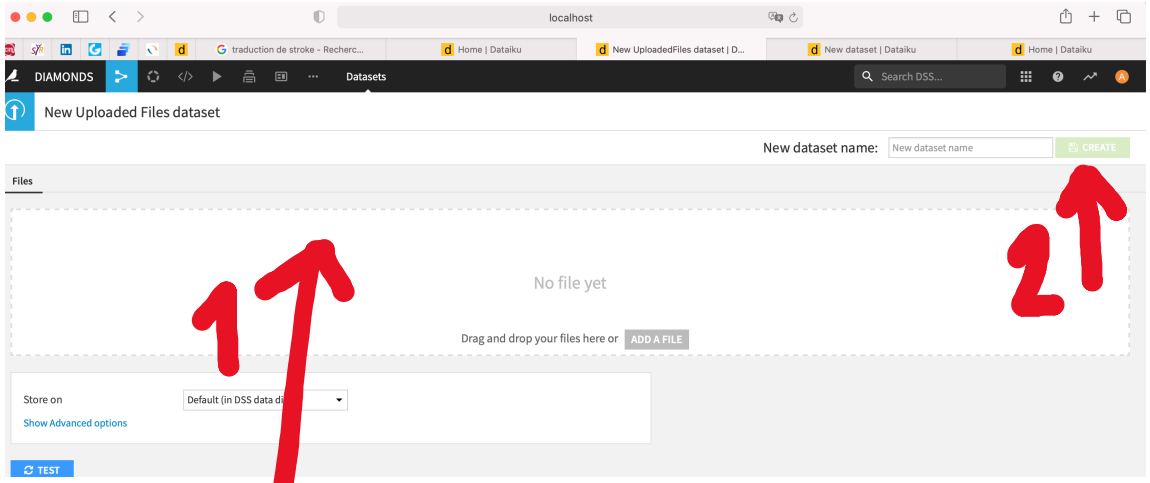

New project
Name
DIAMONDS
Project Key
DIAMONDS
The project key is used to reference datasets between projects. It cannot be changed once the project is created.
CANCEL
CREATE

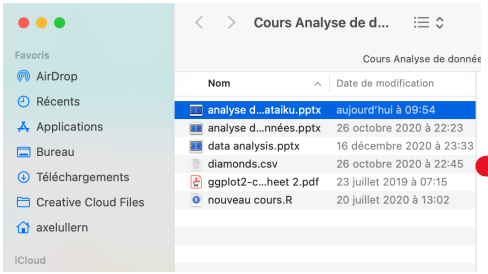



Import your data

# UPLOAD YOUR FILE






1)Drag and Drop of the diamonds file.csv
from your PC folder in the Dataiku Files window
2) Then press create button

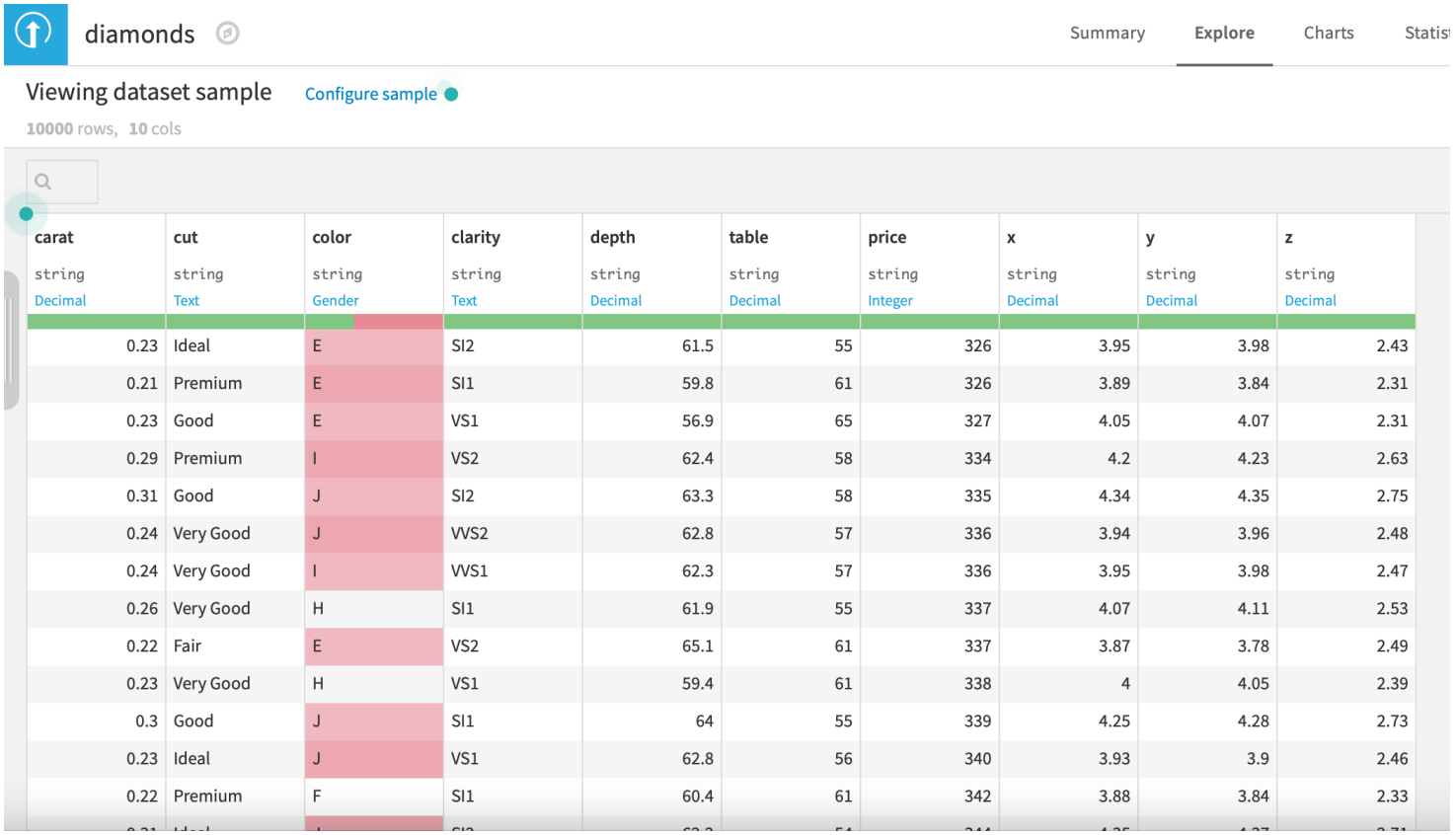
# DATA PREPARATION

DIAMONDS DATASET

DATA DISCOVERY

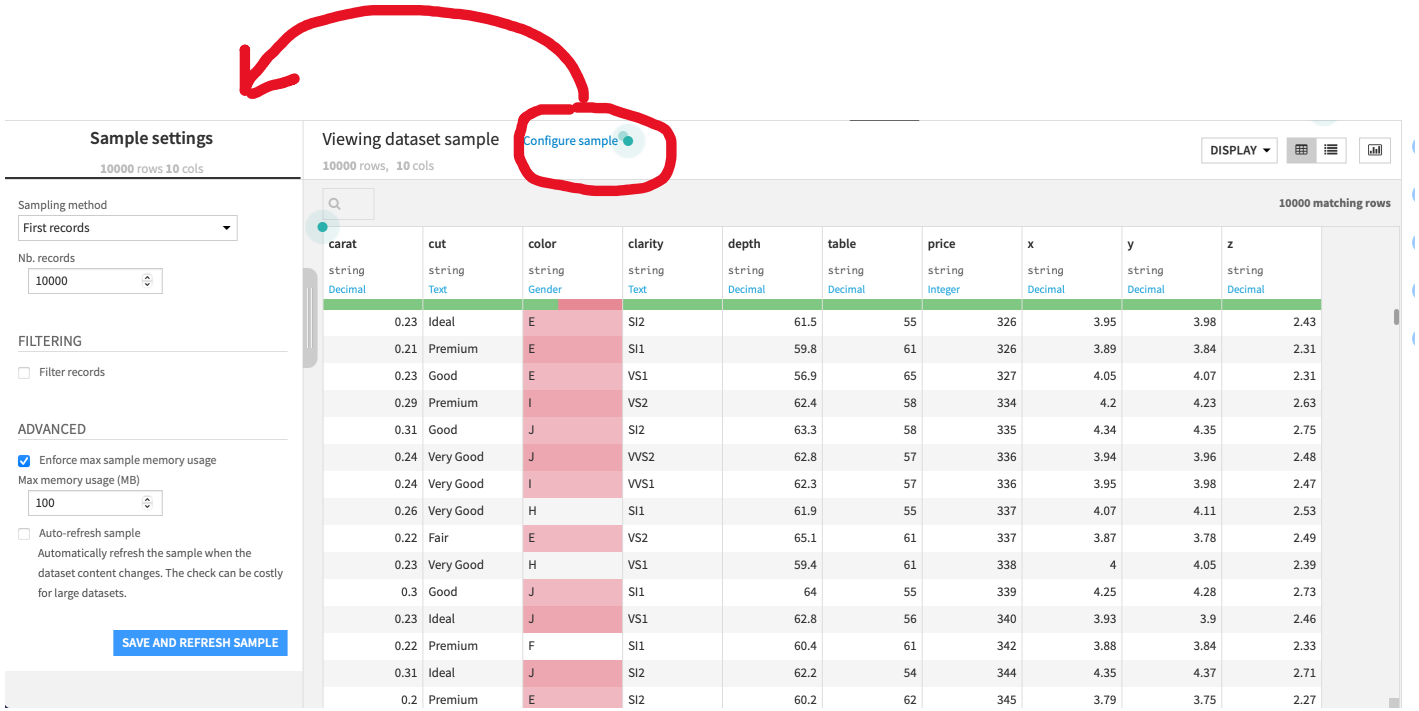
1ST SIMPLE ANALYSES

# Loaded dataset



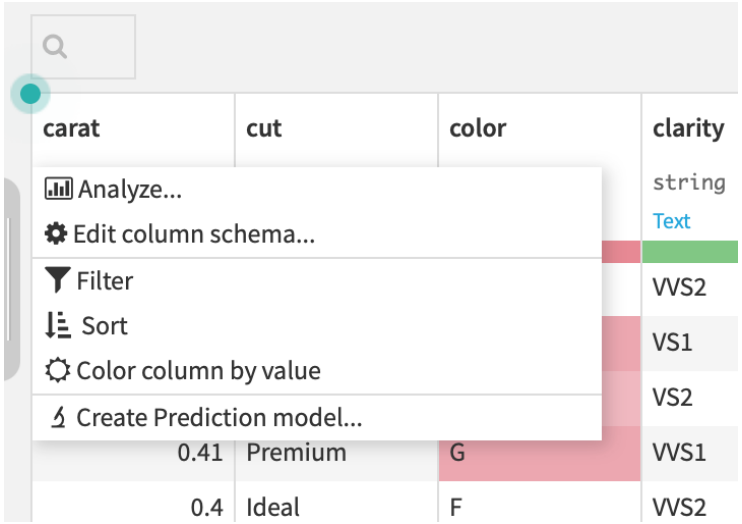

| carat | cut | color | clarity | depth | table | price | x | y | z |
|---|---|---|---|---|---|---|---|---|---|
| string | string | string | string | string | string | string | string | string | string |
| Decimal | Text | Gender | Text | Decimal | Decimal | Integer | Decimal | Decimal | Decimal |
| 0.23 | Ideal | E | SI2 | 61.5 | 55 | 326 | 3.95 | 3.98 | 2.43 |
| 0.21 | Premium | E | SI1 | 59.8 | 61 | 326 | 3.89 | 3.84 | 2.31 |
| 0.23 | Good | E | VS1 | 56.9 | 65 | 327 | 4.05 | 4.07 | 2.31 |
| 0.29 | Premium | I | VS2 | 62.4 | 58 | 334 | 4.2 | 4.23 | 2.63 |
| 0.31 | Good | J | SI2 | 63.3 | 58 | 335 | 4.34 | 4.35 | 2.75 |
| 0.24 | Very Good | J | VVS2 | 62.8 | 57 | 336 | 3.94 | 3.96 | 2.48 |
| 0.24 | Very Good | I | VVS1 | 62.3 | 57 | 336 | 3.95 | 3.98 | 2.47 |
| 0.26 | Very Good | H | SI1 | 61.9 | 55 | 337 | 4.07 | 4.11 | 2.53 |
| 0.22 | Fair | E | VS2 | 65.1 | 61 | 337 | 3.87 | 3.78 | 2.49 |
| 0.23 | Very Good | H | VS1 | 59.4 | 61 | 338 | 4 | 4.05 | 2.39 |
| 0.3 | Good | J | SI1 | 64 | 55 | 339 | 4.25 | 4.28 | 2.73 |
| 0.23 | Ideal | J | VS1 | 62.8 | 56 | 340 | 3.93 | 3.9 | 2.46 |
| 0.22 | Premium | F | SI1 | 60.4 | 61 | 342 | 3.88 | 3.84 | 2.33 |

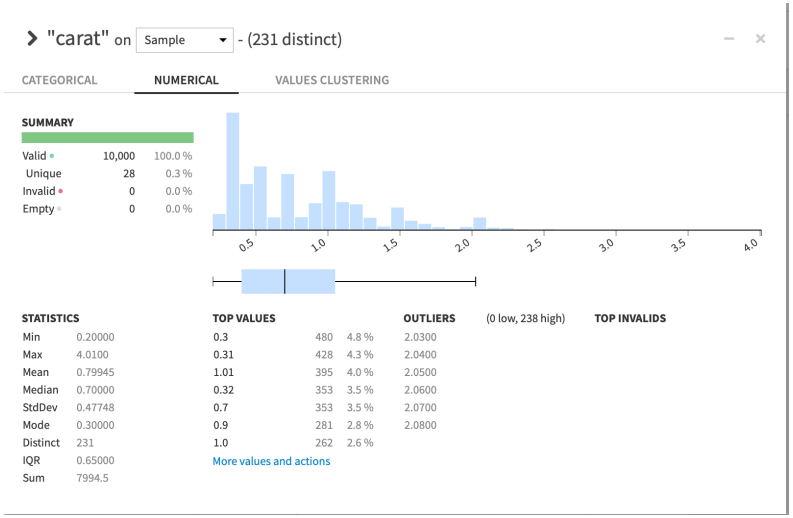
# Configure sampling with configure sample



# ANALYZE EACH COLUMN

- To the right of the name of each column A small triangle
- Click on this triangle and select Analyze



# 1st Analyses

- What is the largest carat?
- What is the average price?
- What percentage of diamonds of the ideal cut?

# Transformations and visualisation

TRANSFORMING DATA WITH RECIPES

VISUALIZE TRANSFORMATIONS WITH FLOW

# RECIPES

**Recipes are used to transform the dataset, these are simple operations such as:**

**Cleaning**

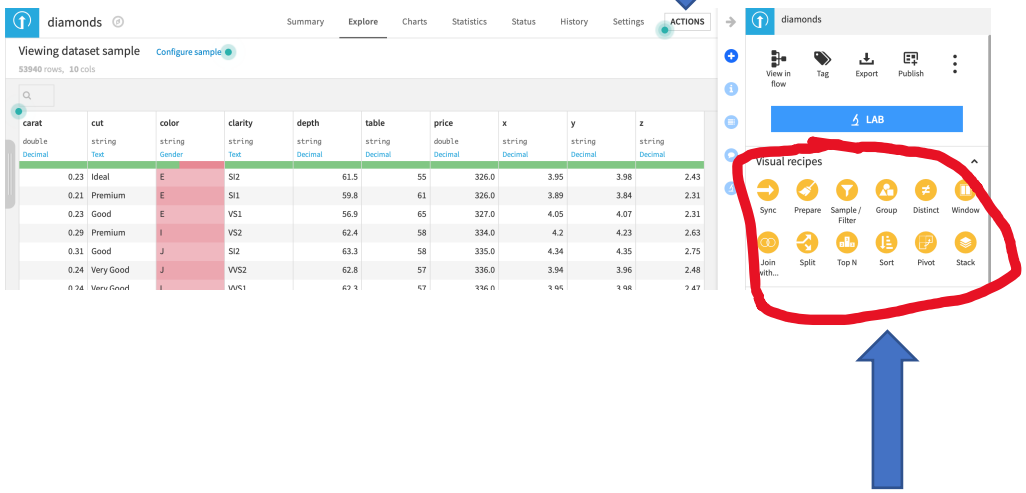
**Renaming**

**Creating a new column,**

**Changing date formats**

**Etc...**

When the dataset
Is open click actions




Visual Recipes are very useful
To prepare to clean up a dataset

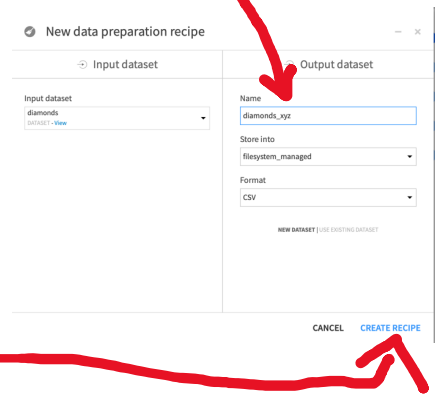
# Dataiku creates a new dataset after running recipe.

- This ensures that the initial data is not overwritten and allow returned to initial dataset  at any time.
- DSS proposes a new name: it is often useful to put a name that corresponds to the transformation performed

# Example: Creating a New Column

In the game diamonds we want to create a volume column whose content is the multiplication of the 3 dimensions x, y, and z (last 3 columns of the dataset)

Step 1: Recipe Prepare Selection

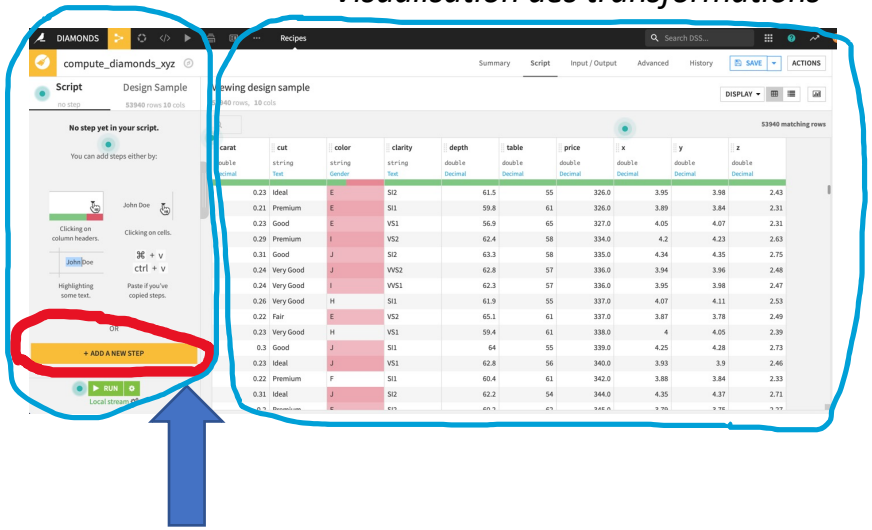

Step 2: Give the new dataset a name
Then create the recipe

# Recipe making

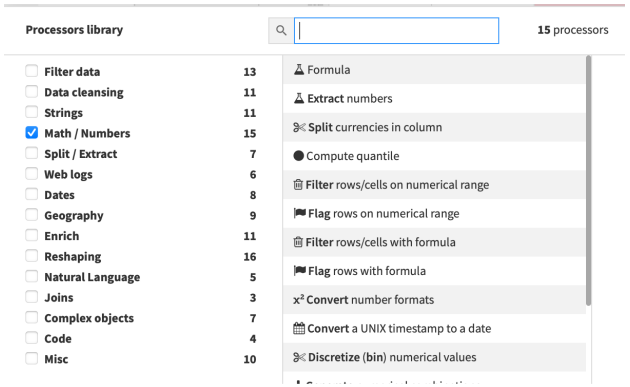
Step 3: Build the recipe script and select the processors

*Script*

*Visualisation des transformations*



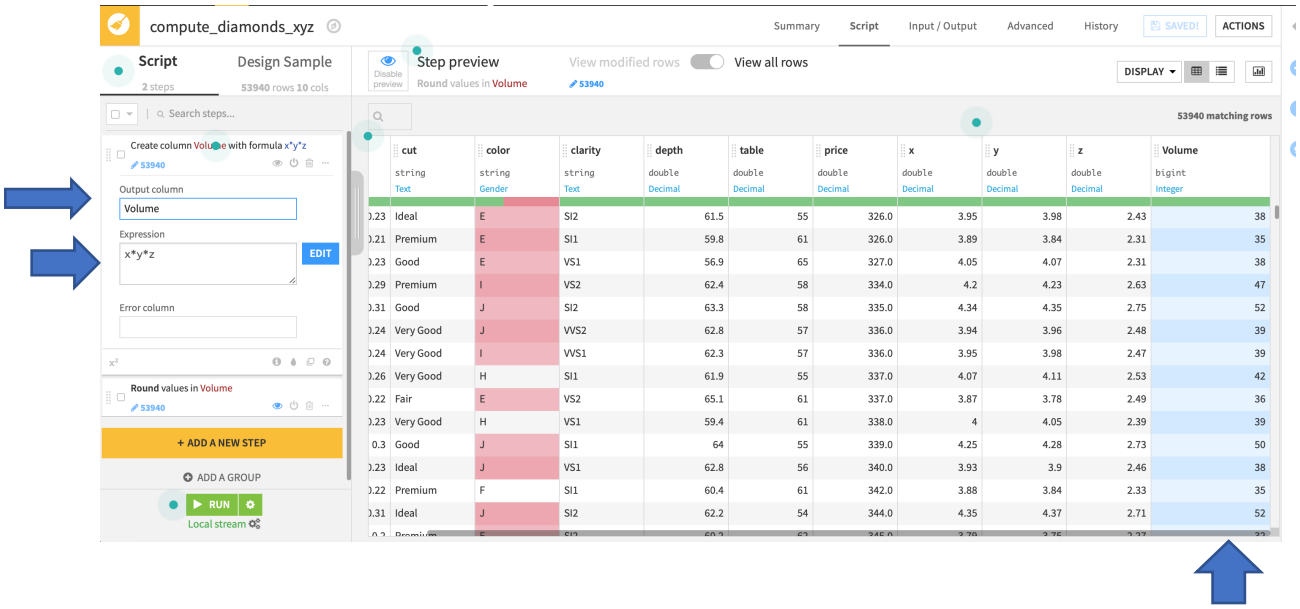

Allows you to add a new
Step in the script and open the window for choosing processors

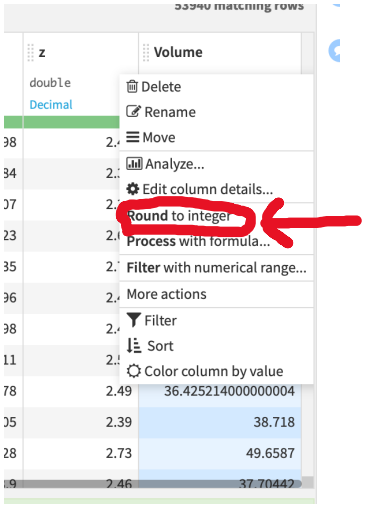



Here the chosen processor is Maths/Numbers because we want to Create a new column Calculated from the others existing columns

# Creating the new column in the recipe

Step 4: we name the new volume column and indicate the calculation formula: Volume=x*y*z (x, y and z are the names of the last 3 columns of the dataset)



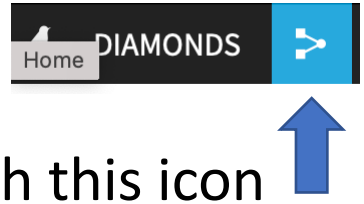

Allows rounding Decimal digits

The new column is created

# Last step: Execute the recipe

- Click on Run and wait for the "Job Succeeded" message
- WARNING: AS LONG AS THIS OPERATION HAS NOT BEEN DONE THE NEW DATASET IS NOT CREATED!

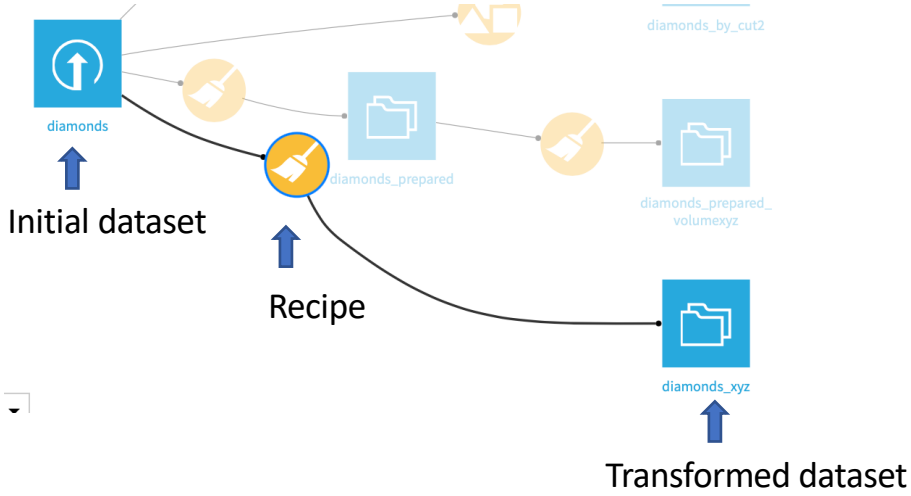
# FLOW



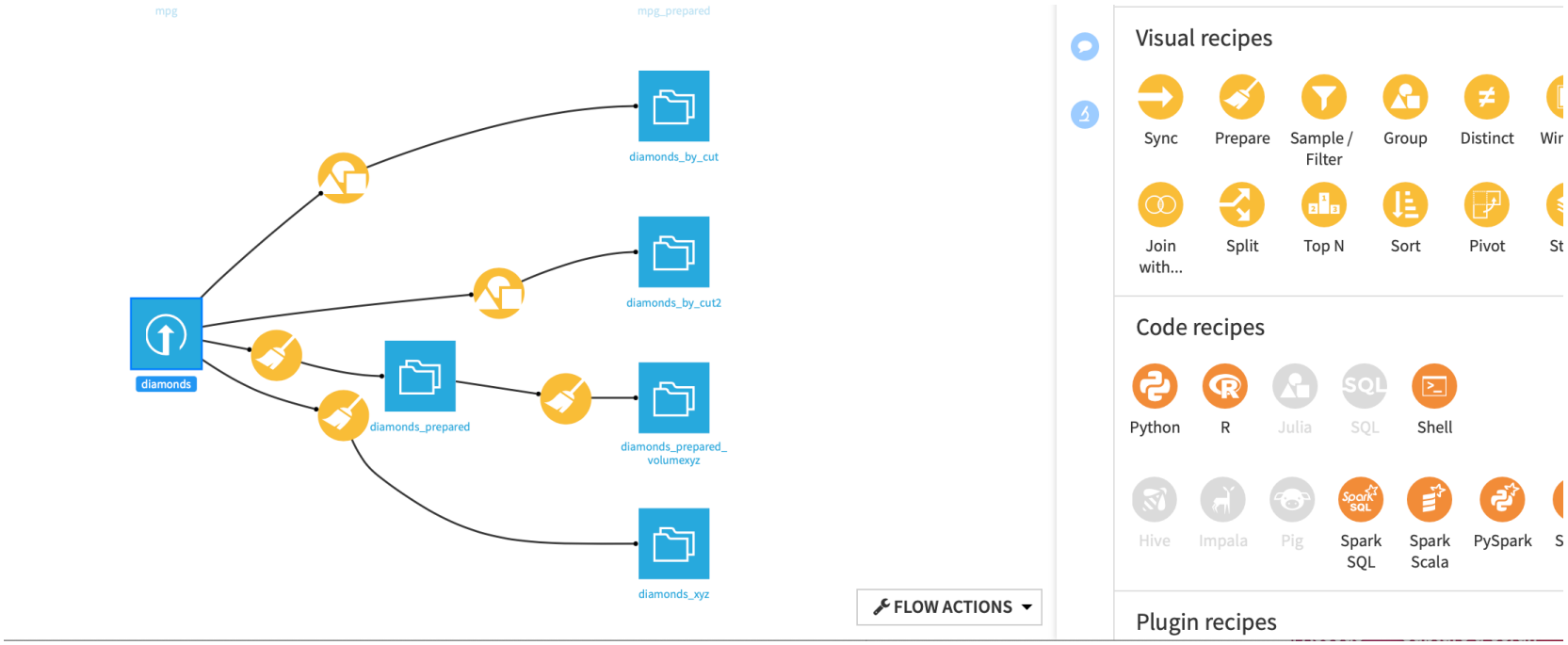

It is accessible at any time with this icon

It shows datasets and recipes

It allows you to see all the stages of a project



# Flow example



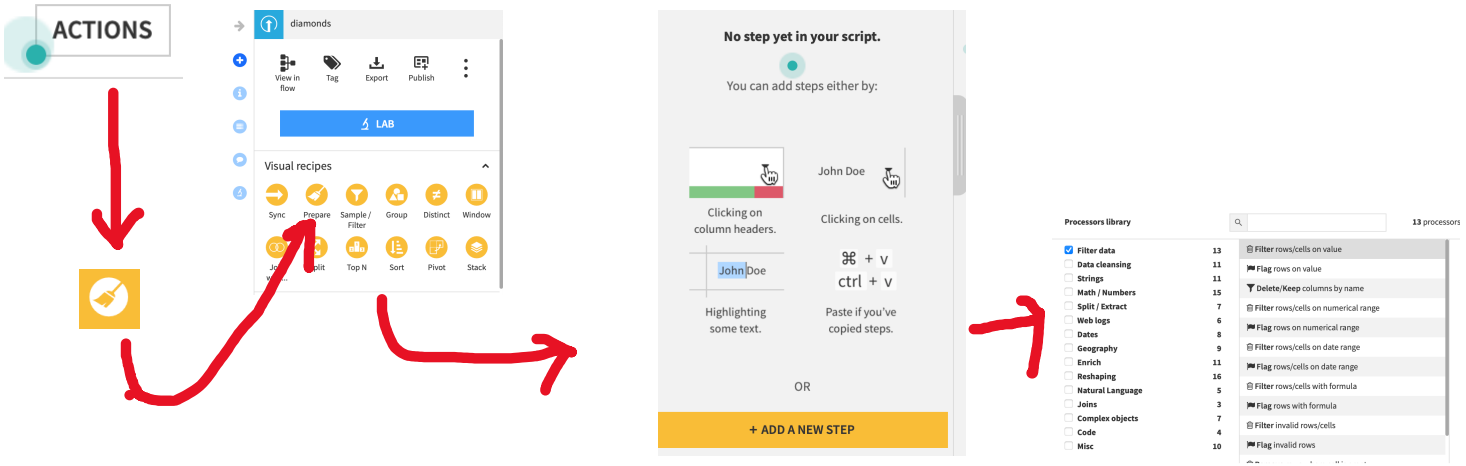

From the dataset diamonds we created 5 new datasets and 5 recipes (2 groups and 3 preparations)

# ANOTHER RECIPE: FILTERING

Objective: This is to filter the rows of the table on the basis of criteria (e.g. prices < $ 1000, carats >3)

With Dataiku 2 ways to filter:
Filtering directly on the column: no modification of the data table
Filtering with recipe: Creating a new filtered dataset

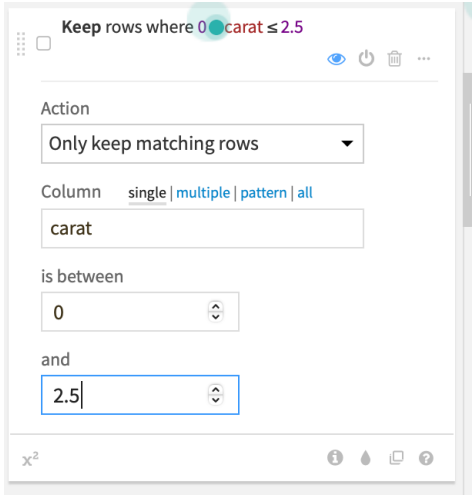
Filtering with recipe:



# Practice

Create a dataset from the diamond set containing the carats<2.5

Here's what to
Put in recipe
Filtering

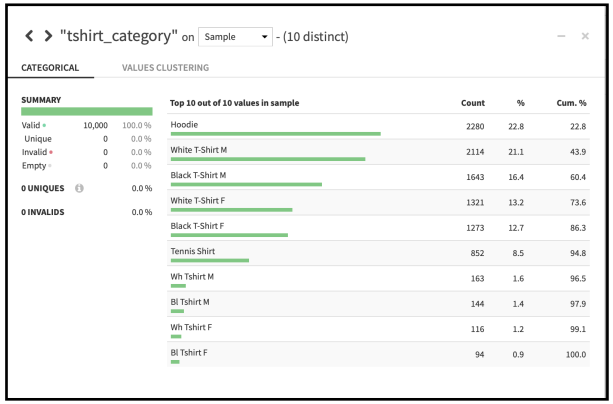



*You will examine the transformed dataset to check in the carat column that all carat are now >2.5*
*then run the recipe and observe the Flow*

# DATA PREPARATION

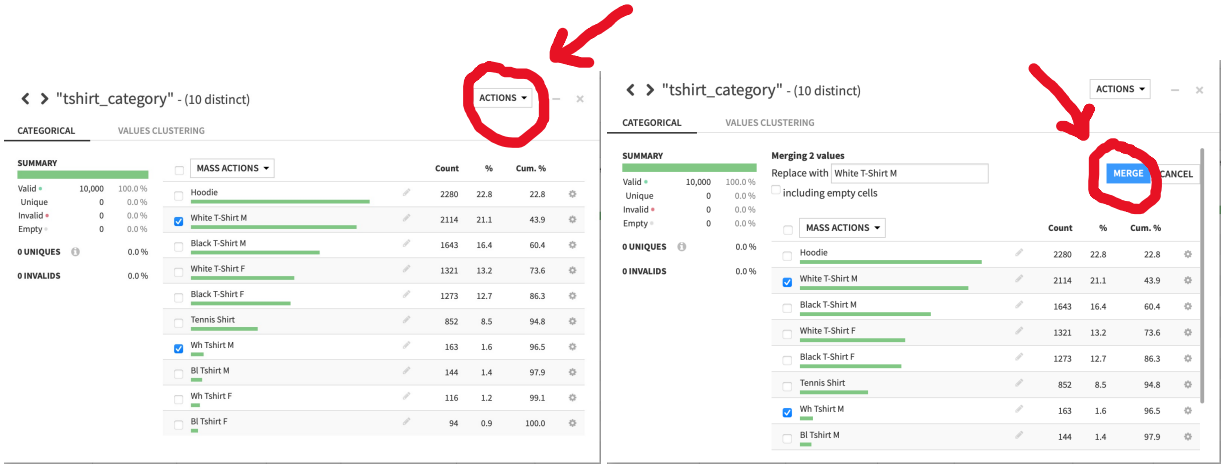
- Many datasets have errors in the coding (labels) of the data
- that may distorting the analysis, Dataiku will very easily allow us to change these codifications
- Ex: load into Dataiku the dataset "orders" which includes sales of T-Shirt.

*By clicking on the triangle*
*Next to the column*
*tshirt_category then Analyze*
*The Marketing Product Manager*
*Realizes that many*
*Codification errors have been made:*
*White T shirt M and Wh Tshirt M are the same*
*Products and must be grouped in*
*The same category.*



# Merge product categories into 1 single

- Create a prepare recipe and then directly in the recipe click on the t shirt category column
- And select the two categories to merge then click on ACTIONS then MASS ACTIONS, accept or change the name of the merged category then click MERGE



# Exercise: Finish the recipe

- Merge the other categories of erroneous Tshirt with the right categories in the same way
- Then do RUN when the recipe is ready
- Examine the flow
- Open the modified dataset
- Verify that all categories are correct

# ANALYSIS AND INSIGHTS

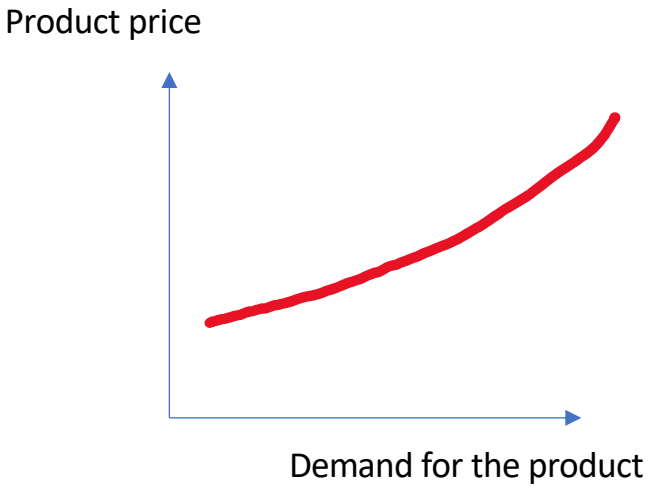
Choosing data to visualise

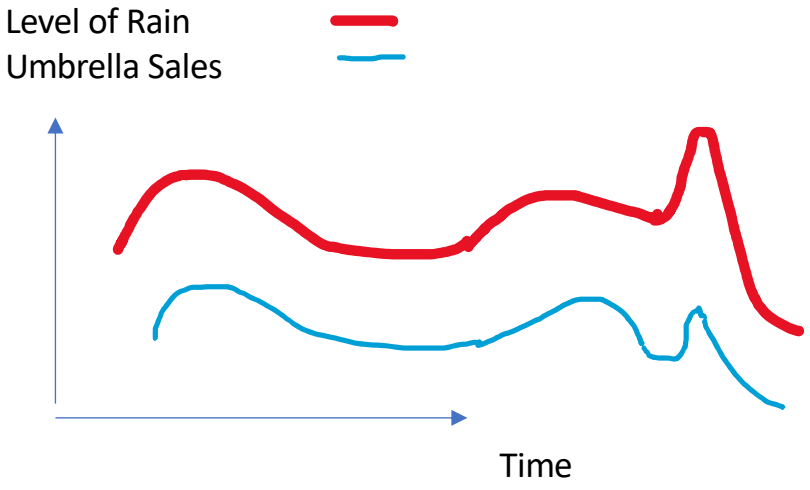
Creating graphs

Get insights from graphs

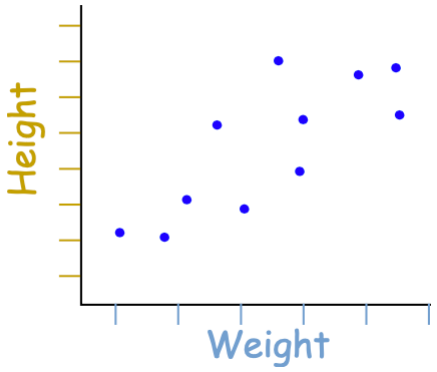
# What do we see on a graph ?

This makes possible to see the **link (or influence)** between two variables: the one that is on the abscissa and the one that is in ordinates.
The correlation between several variables can also be studied by comparing their respective evolution

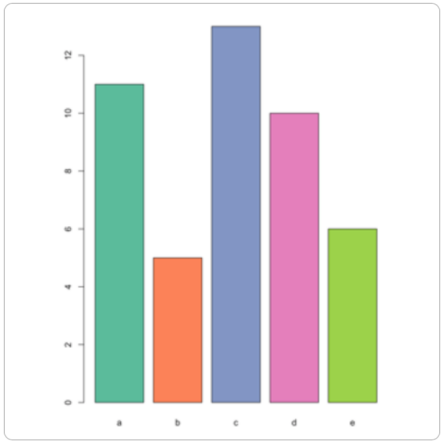


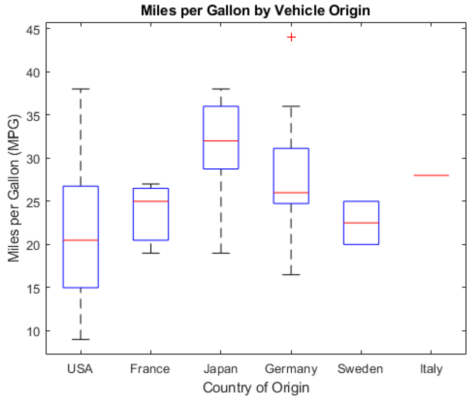
# Main graphs type




Scatterplot or dot diagram
**The two variables in X and Y are Continues**
*Allows you to see the possible link between The two variables*

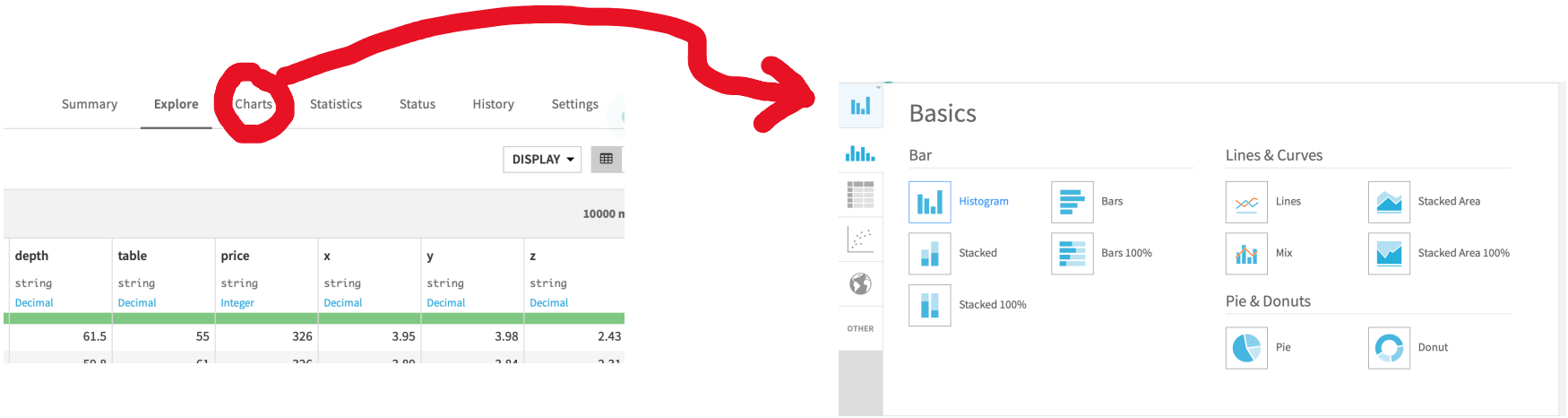


Bar plot or bar chart
**The variable in X is categorical**
**No Y** variable: the height of the bar
Corresponds to the quantity of the variable X
In each category
*Allows you to compare the importance of categories*




Boxplot or Whisker Box
**The variable in X is categorical**
**The variable in Y is continuous**
*Allows you to compare different categories*
*In X between them according to the variable In Y*

# Graphs with Dataiku



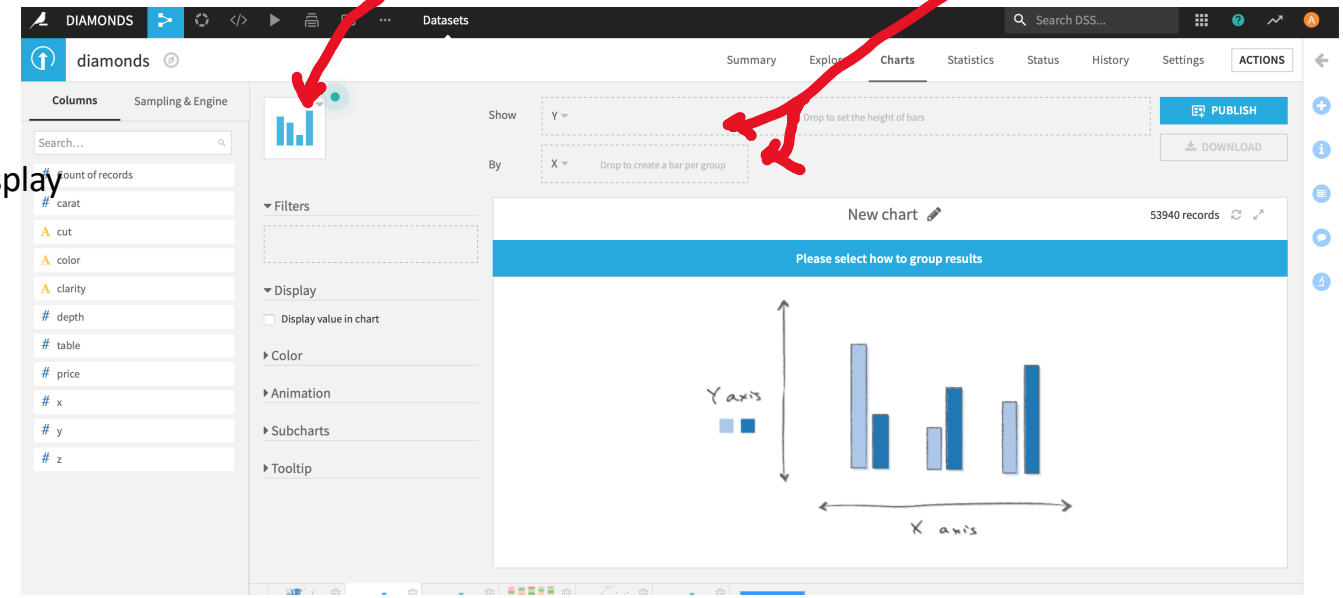

Click on "Charts" then choose the type of graph and build the graph from the data

# Building Graph

1 Choose graph type

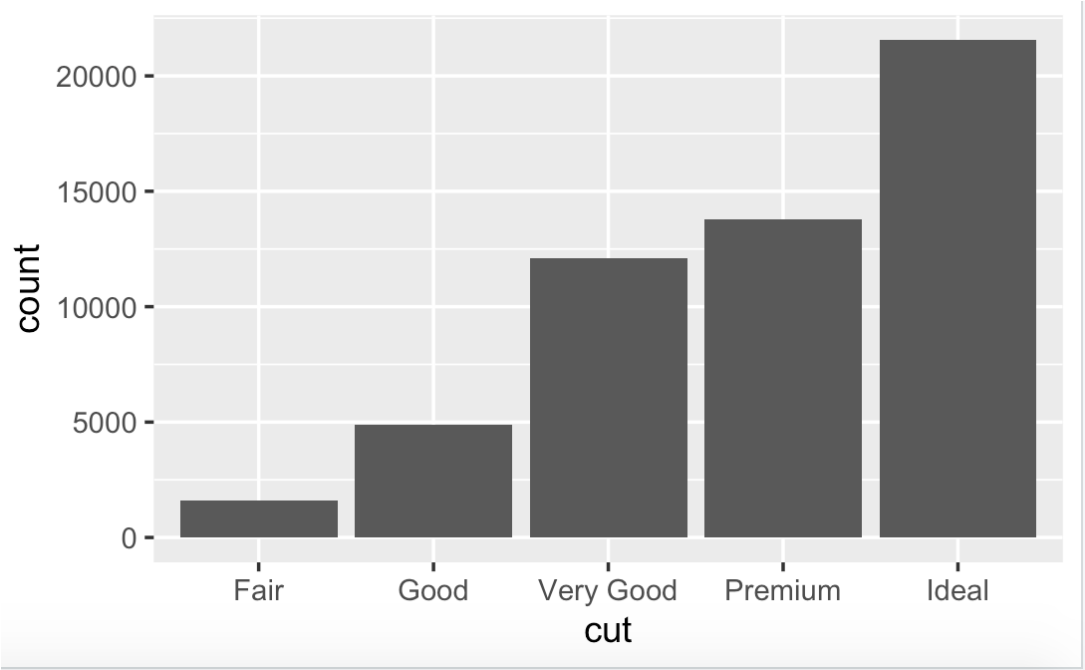
3 drag and drop on X and Y

2 Sélect variables to display in X the Y



# BAR GRAPH

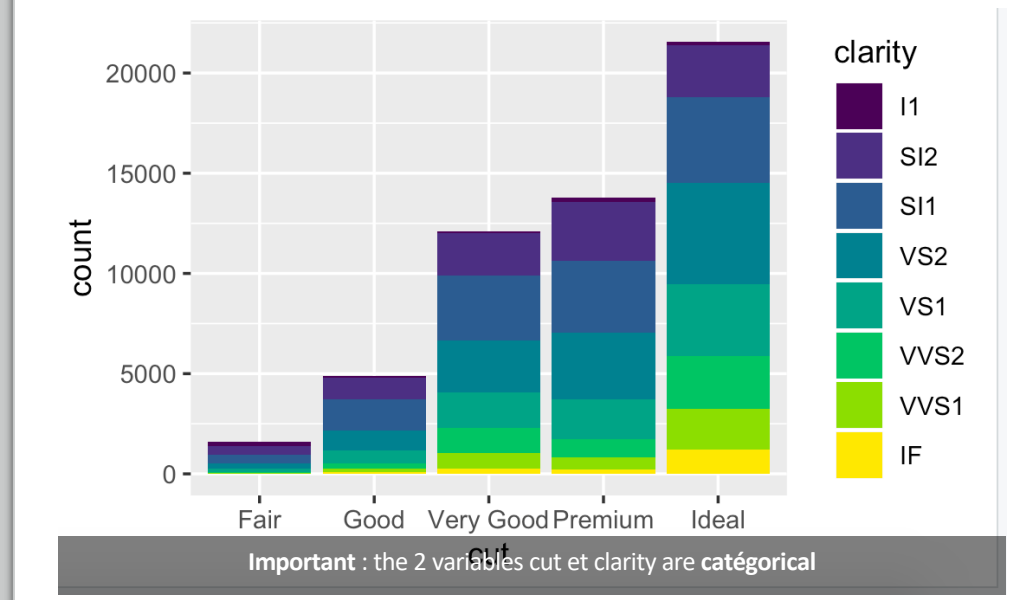
This graph is adapted to categorical (or discontinuous) variables In X the categories of the variable , in Y the number of occurrences of the variable in the dataset




It is easy to see how much diamonds of each category in the dataset

# Bars stacked

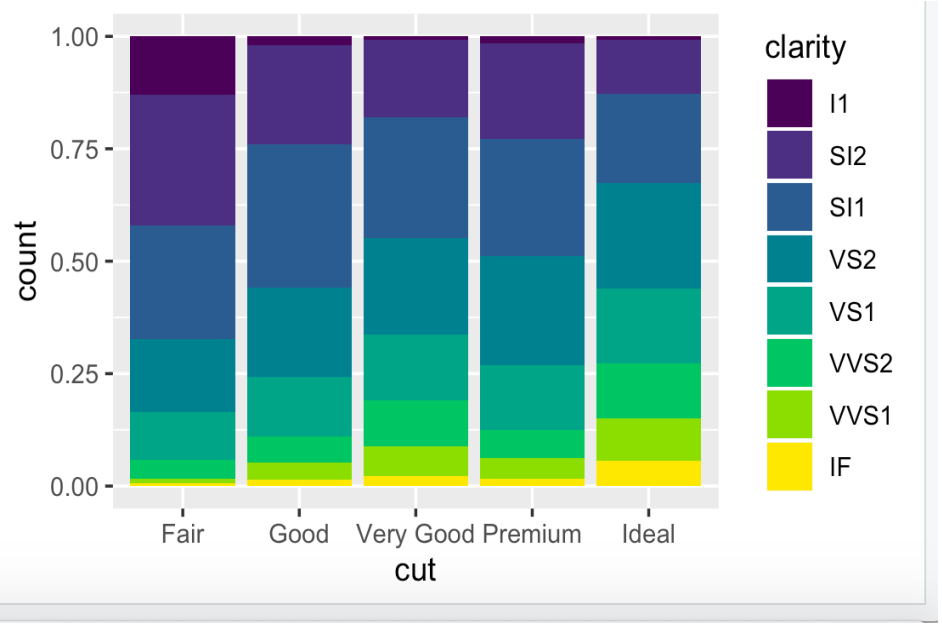
- The bars are stacked with a second variable (clarity )
- We can see how much clarities in the different cut (1st variable)



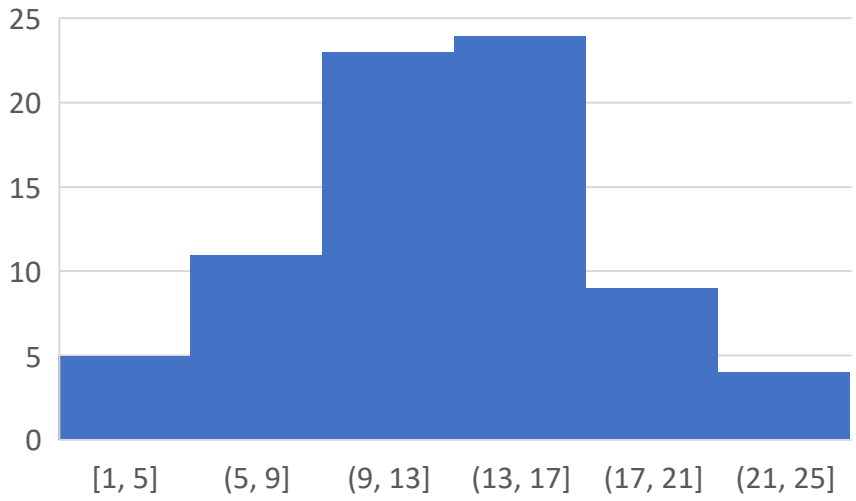

**Important** : the 2 variables cut et clarity are **catégorical**

# Bars Stacked 100%

It is easier to compare the inside of the bars if they have the same height



# HISTOGRAM




The histogram allows you to cut
A continuous variable in "slices"
Called "bin"
Each slice behaves like
one of the categories
of a categorical variable

For the dataset diamonds this type of
Graph adapts well to the
Carat variable on the abscissa
And counting the number of diamonds
per carat slice

# OBJECTIVE: to determine which factor(s) most influence the price of diamonds

Cut ? Carat? Color ? ……

# Exercise : Display link between 2 then 3 variables Try to determine what is influencing price

1. You want to see if there is an influence of cut on price :
    1. Which is the nature of those 2 variables ?
    2. Which graph will you choose ?
    3. Build the graph in dataiku
    4. Then try to interpret it ? Does the result confirm Expectations ?
2. You want to compare the colors inside the different cut
    1. Which graph ?
    2. Build the graph
    3. What do you observe ?
3. Same for clarity inside cut
4. Influence of carat on price ? (same questions )

***Price**: price in US dollars (\$326--\$18,823)*
***Carat**: weight of the diamond (0.2--5.01)*
***Cut**: quality of the cut (Fair, Good, Very Good, Premium, Ideal)*
***Color**: diamond colour, from D (best) to J (worst)*
***Clarity**: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))*

# Exercice : Filter and analyse deeper

1. On former exercise you made a graph with carats on the abscissa and prices on ordinates What is strange with this graph ?
1. On the same graph Color the dots according to cut  Are you satisfied with the result ?
2. Filter the diamonds game by selecting the large carats > 2.4
3. Then make a graph carat / price with colored points based on the cut

Let's talk business!

# The company Telco is very concerned about the lack of loyalty of its customers and would like to understand the reasons for the loss of its customers (churn)

As an external consultant,
You will analyze the situation
of this company
will make recommendations

- Load Telco Churn datasetHow many churn vs no churn clients? (Churn rate)
- Does the client's gender have an influence on churn? Base your answer on a graph
- Is the InternetService, TechSupport, ... have an influence on the Churn? Base your answer on a graph
- Do TotalCharges have an influence on churn? Create the appropriate chart that will support your answer
- What is tenure? How is it related to Churn?

**In** conclusion: What recommendations do you make to the Telecom operator to better retain its customers ?

- Churn =Yes the customer has broken his contract or no longer buys (lost customer) Churn= No the customer is kept (loyalty
- )

# DATA PREPARATION AND TRANSFORMATION

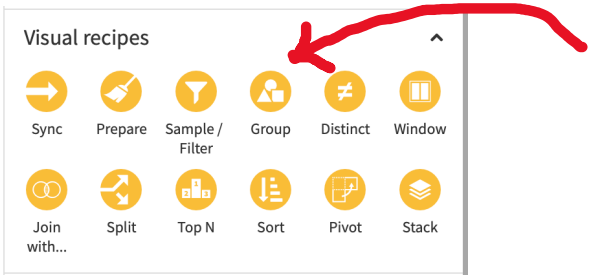
# Aggregation (group by)

**Theory: Aggregation** consists of concentrating the dataset according to the categories of a variable
We then have one line per category
**Example:** here are the grades of 3 high school students, there are several lines for the same high school student because each has several grades
Aggregate consists in calculating the average (or the sum) of the grades of a student
we then would have a new table
With one row per student and its average grade: the new table is more synthetic

| STUDENT | SUBJECT | GRADE |
|---|---|---|
| ARTHUR | MATHS | 12 |
| INES | PHYSICS | 15 |
| SEBASTIEN | FRENCH | 8 |
| ARTHUR | FRENCH | 11 |
| INES | MATHS | 11 |
| SEBASTIEN | PHYSICS | 16 |

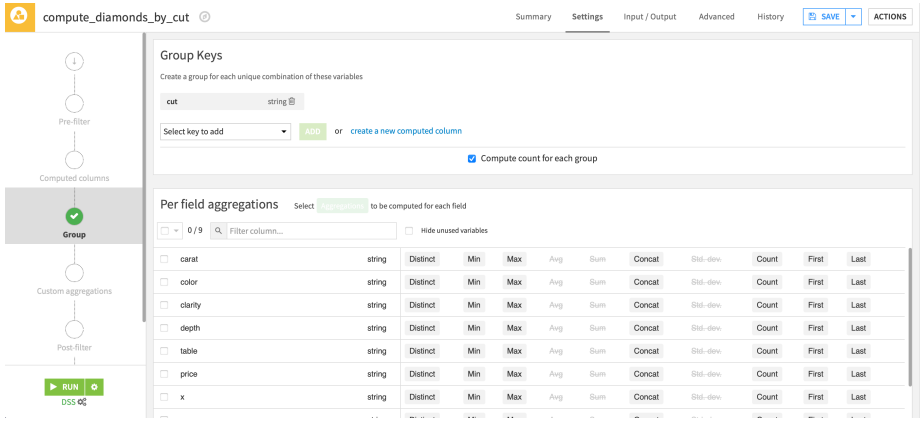
*Here the agregation key*
*Is the student :*
*We would like to group*
*by student the table*
*And calculate average*
*grade per student*

# Aggregation with Dataiku

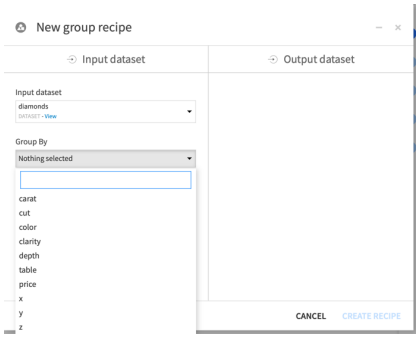



1)Choose recipe Group

3) Specify which calculations have to be made in the final table




2)select aggregation Key



# Application :Group by

Group diamonds per cut

Calculate average price per cut

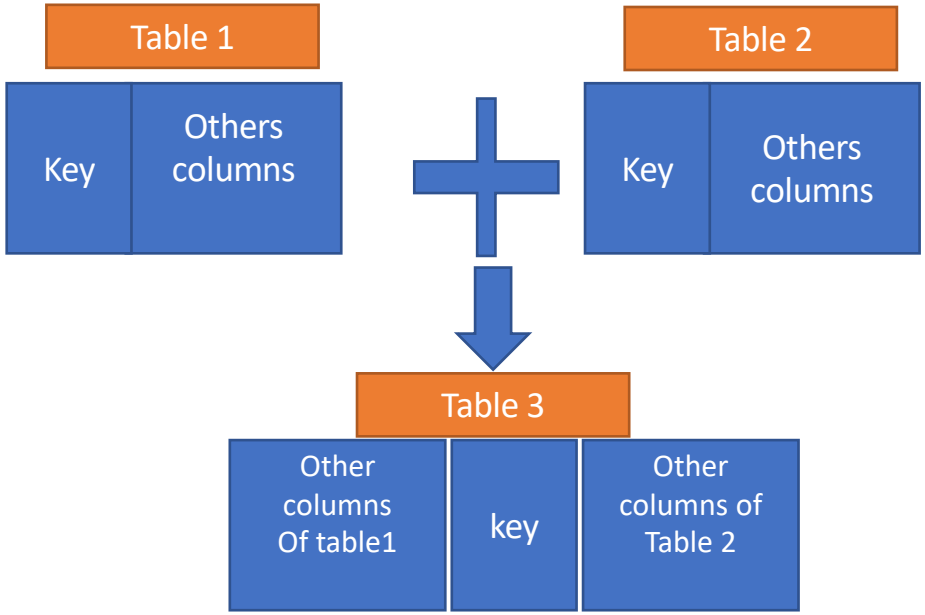
# JOINING TABLES

IT IS VERY COMMON AND USEFUL TO JOIN TABLES

IN ORDER TO GET A NEW TABLE CONTAINING LOTS OF INFORMATIONS

# ENRICH DATA WITH JOIN BETWEEN 2 TABLES

Two tables can only be "joined" if they have
A common column (variable), this variable is called "key"

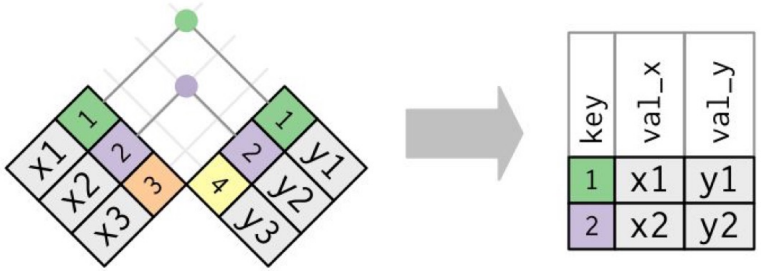



We get in the new table 3 all informations we had in both table 1 and 2
(the key is not repeated twice !)

*The column key in table 1 and the column key in table 2*
*may have different names provided*
*The information in the 2 columns are the same*

# Several kind of joins

**Inner join** =Only rows that have the same value in the Key column are retained in the new table




**Inner Join**

Rows 3 and 4 are dropped in the join
The result of the join is a table with
A minimum number of rows

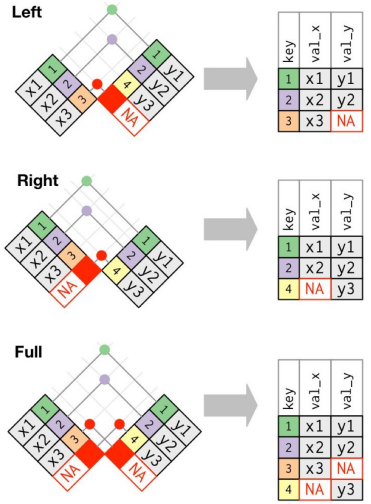
# OUTER_JOIN

## Outer Joins

An inner join keeps observations that appear in both tables. An *outer join* keeps observations that appear in at least one of the tables. There are three types of outer joins:

- A *left join* keeps all observations in `x`.
- A *right join* keeps all observations in `y`.
- A *full join* keeps all observations in `x` and `y`.

These joins work by adding an additional "virtual" observation to each table. This observation has a key that always matches (if no other key matches), and a value filled with `NA`.



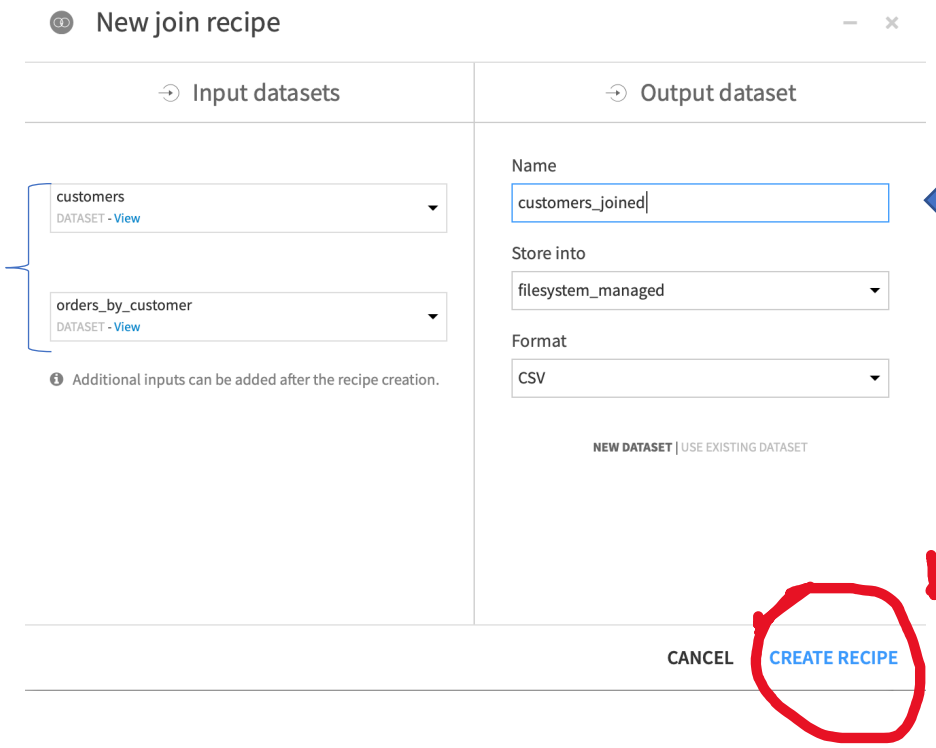
# JOIN WITH DATAIKU

**Datasets: orders.csv and customers.csv**

1) Pre-join data preparation: Aggregate the orders.csv table so that there is only one row per customer
2)Select the two datasets to join  from the flow (right click on each)
3)Select the join recipe
4) Give the name to the new merged table (or keep the proposed name)
5) Check Dataiku has found the key
6) Change the key (if needed) by pressing the equal sign)
7) Click on Create

# Building the join recipe

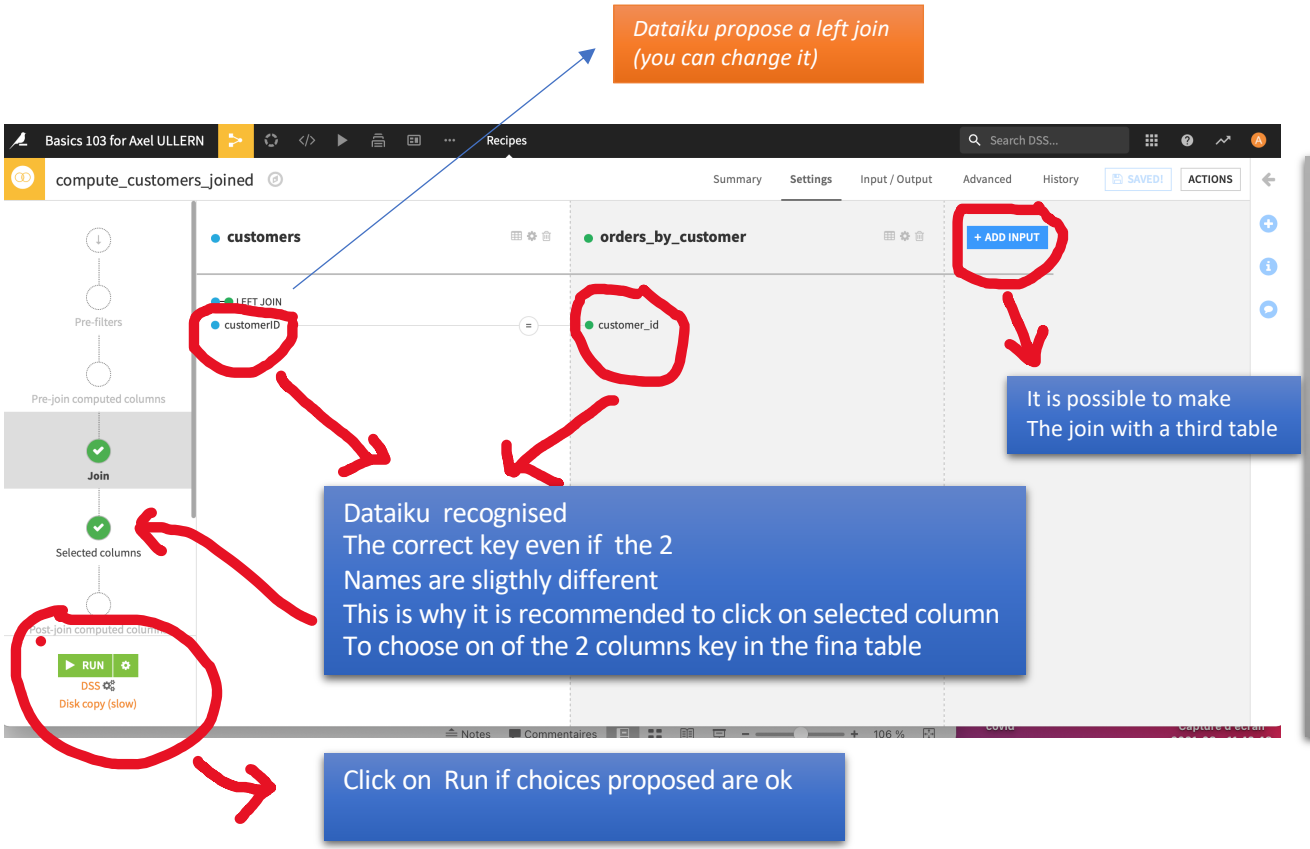
Datasets names to be joined




Name of new Dataset

Click here

# The join recipe

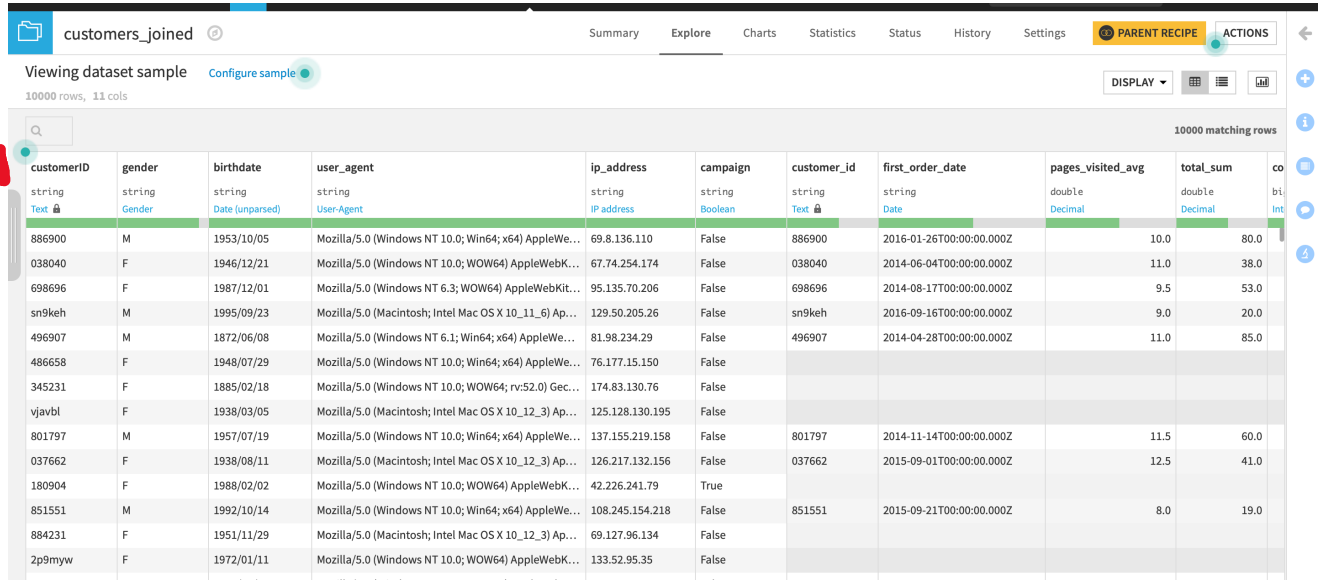



***Attention :***
***If the join key***
***Has a different wording in both datasets***
***Then you have to click on the sign =***
***To specify the key***
***and select***
***Manually***
***The 2 columns***

# The final result

key

Check all columns from original datasets are present




| customerID | gender | birthdate | user_agent | ip_address | campaign | customer_id | first_order_date | pages_visited_avg | total_sum | co |
|---|---|---|---|---|---|---|---|---|---|---|
| string | string | string | string | string | string | string | string | double | double | bi |
| Text | Gender | Date (unparsed) | User-Agent | IP address | Boolean | Text | Date | Decimal | Decimal | Int |
| 886900 | M | 1953/10/05 | Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWe… | 69.8.136.110 | False | 886900 | 2016-01-26T00:00:00.000Z | 10.0 | 80.0 | |
| 038040 | F | 1946/12/21 | Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebK… | 67.74.254.174 | False | 038040 | 2014-06-04T00:00:00.000Z | 11.0 | 38.0 | |
| 698696 | F | 1987/12/01 | Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit… | 95.135.70.206 | False | 698696 | 2014-08-17T00:00:00.000Z | 9.5 | 53.0 | |
| sn9keh | M | 1995/09/23 | Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_6) Ap… | 129.50.205.26 | False | sn9keh | 2016-09-16T00:00:00.000Z | 9.0 | 20.0 | |
| 496907 | M | 1872/06/08 | Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWe… | 81.98.234.29 | False | 496907 | 2014-04-28T00:00:00.000Z | 11.0 | 85.0 | |
| 486658 | F | 1948/07/29 | Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWe… | 76.177.15.150 | False | | | | | |
| 345231 | F | 1885/02/18 | Mozilla/5.0 (Windows NT 10.0; WOW64; rv:52.0) Gec… | 174.83.130.76 | False | | | | | |
| vjavbl | F | 1938/03/05 | Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_3) Ap… | 125.128.130.195 | False | | | | | |
| 801797 | M | 1957/07/19 | Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWe… | 137.155.219.158 | False | 801797 | 2014-11-14T00:00:00.000Z | 11.5 | 60.0 | |
| 037662 | F | 1938/08/11 | Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_3) Ap… | 126.217.132.156 | False | 037662 | 2015-09-01T00:00:00.000Z | 12.5 | 41.0 | |
| 180904 | F | 1988/02/02 | Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebK… | 42.226.241.79 | True | | | | | |
| 851551 | M | 1992/10/14 | Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWe… | 108.245.154.218 | False | 851551 | 2015-09-21T00:00:00.000Z | 8.0 | 19.0 | |
| 884231 | F | 1951/11/29 | Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_3) Ap… | 69.127.96.134 | False | | | | | |
| 2p9myw | F | 1972/01/11 | Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebK… | 133.52.95.35 | False | | | | | |

# A REAL CASE : Employee Attrition

CREATE A NEW PROJECT IN DATAIKU

IMPORT DATASET « HR-EMPLOYEE-ATTRITION »

# The Turnover seems too high: can you find why ?

You are commissioned by the HR Director as a specialized consulting firm to help this company.

You will analyze the data and extract as much information as possible to understand the situation and determine potential causes of turnover

Then you will formulate your recommendations that will be presented to the Executive Committee

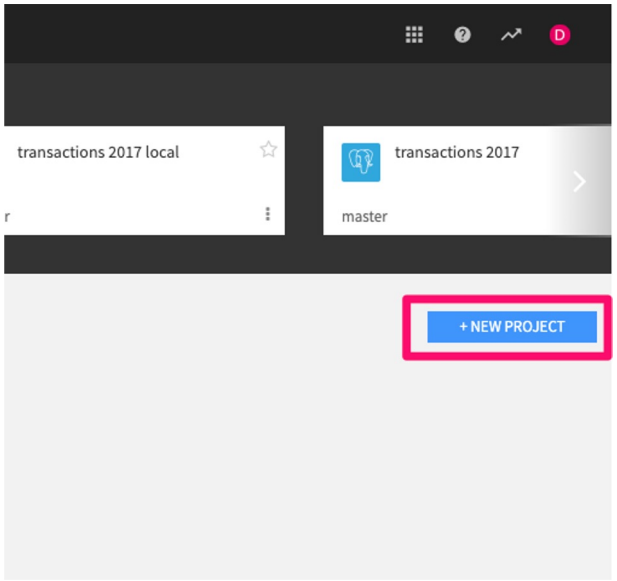
# COMPLETE ANALYSIS OF CREDIT CARD FRAUD

# Agenda

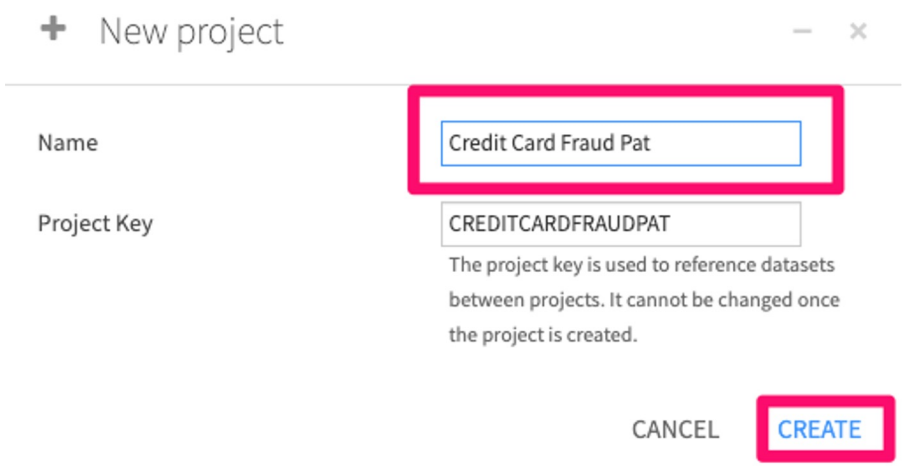
Discovery
Hands-on Exercise

1. Background
2. DSS Login and Set up Account
3. **Creating Projects and Connecting to Data**
4. Stacking Datasets
5. Joins
6. Computation Engines
7. Data Cleaning [1]
8. Charts [1]
9. Splitting Data
10. Machine Learning
11. Filtering Data
12. Data Cleaning [2]
13. Group by Aggregations
14. Charts [2]

# Create a new project




Name your project “Credit Card Fraud” + your name



## Data Sources Overview: 4 Datasets

**transactions_2017** and **transactions_2018**

- consumer credit card transactions

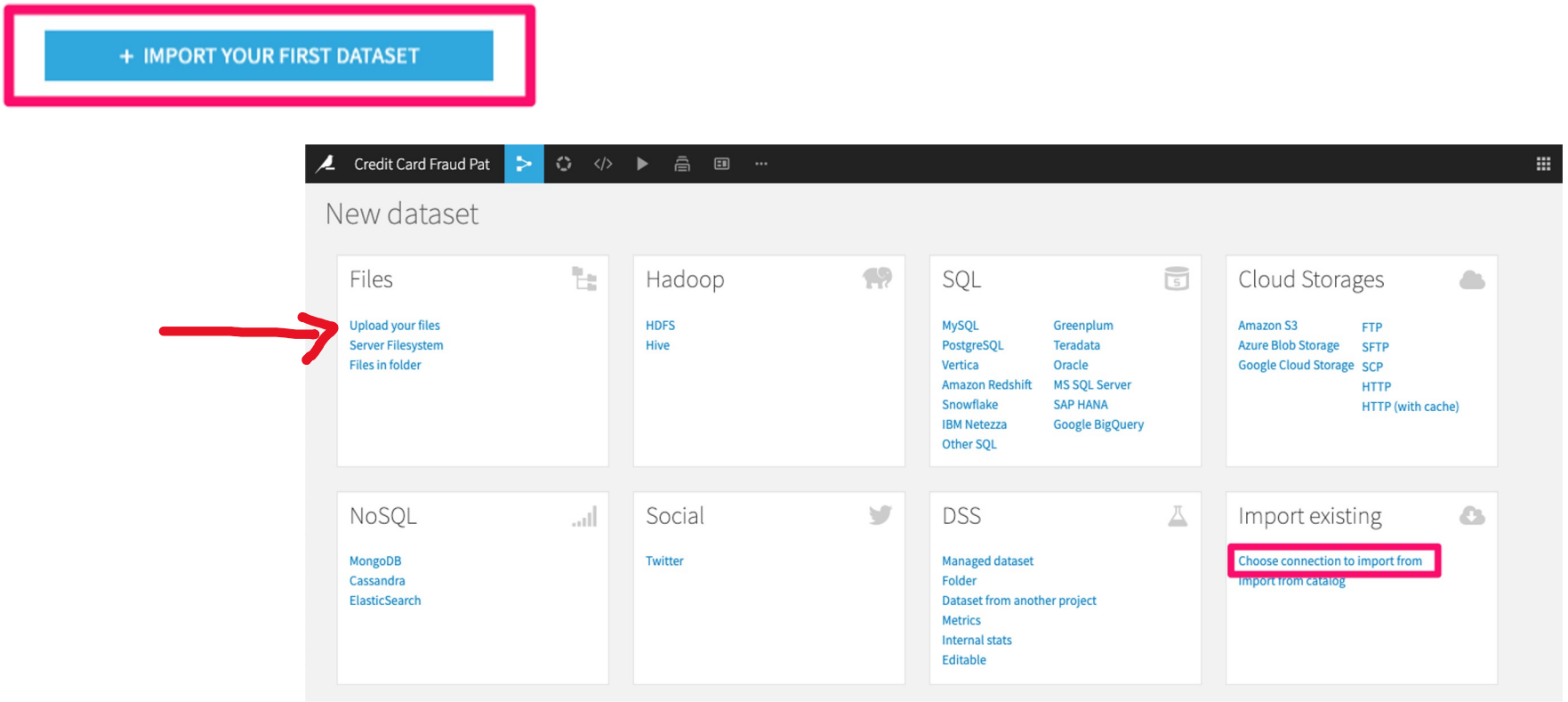
**cardholder_info**

- credit score for each cardholder

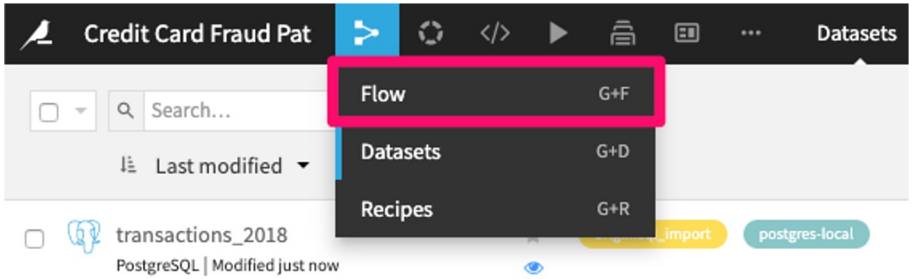
**merchant_info**

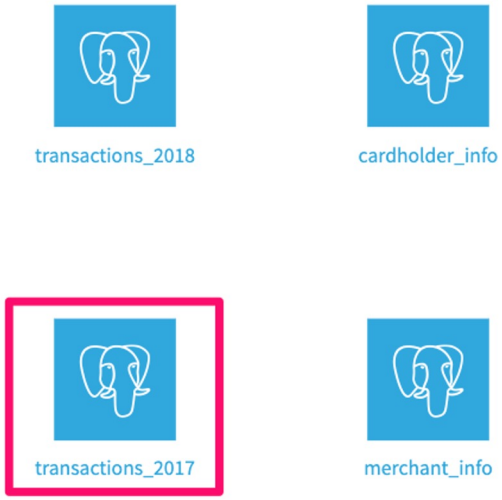
- Lookup table for merchants

# Import data


+ IMPORT YOUR FIRST DATASET
Credit Card Fraud Pat
New dataset
Files
Upload your files
Server Filesystem
Files in folder
Hadoop
HDFS
Hive
SQL
MySQL
PostgreSQL
Vertica
Amazon Redshift
Snowflake
IBM Netezza
Other SQL
Greenplum
Teradata
Oracle
MS SQL Server
SAP HANA
Google BigQuery
Cloud Storages
Amazon S3
Azure Blob Storage
Google Cloud Storage
FTP
SFTP
SCP
HTTP
HTTP (with cache)
NoSQL
MongoDB
Cassandra
ElasticSearch
Social
Twitter
DSS
Managed dataset
Folder
Dataset from another project
Metrics
Internal stats
Editable
Import existing
Choose connection to import from
Import from catalog

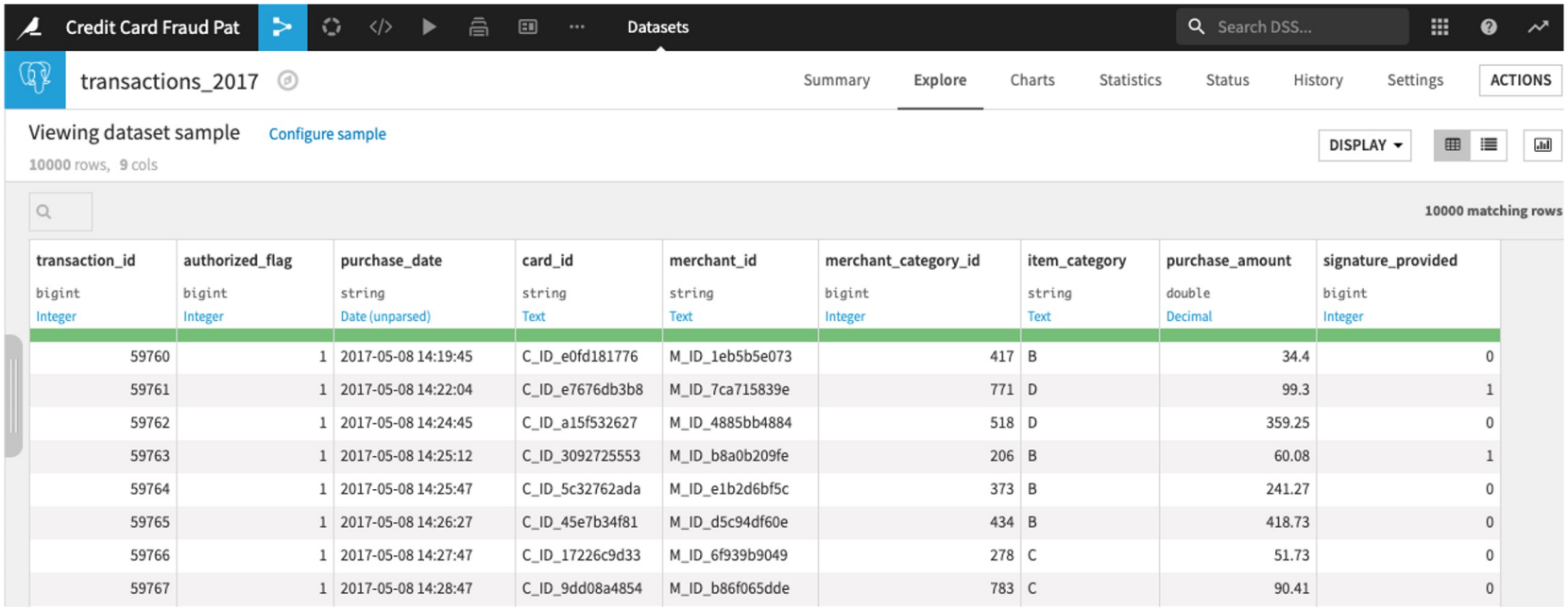
# You should now have 4 datasets in your flow




Double click on a dataset to view it

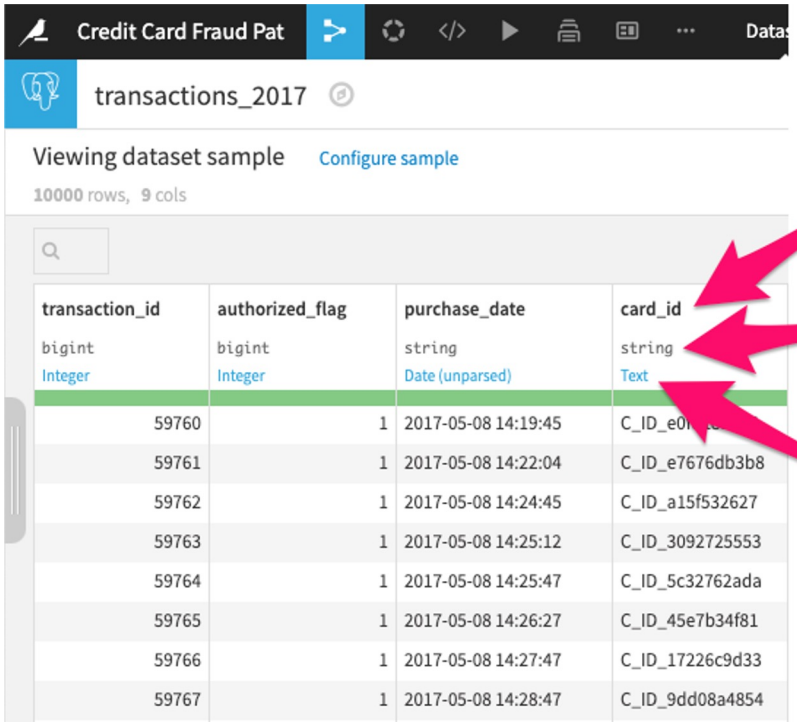
# Explore the transaction data




| transaction_id | authorized_flag | purchase_date | card_id | merchant_id | merchant_category_id | item_category | purchase_amount | signature_provided |
|---|---|---|---|---|---|---|---|---|
| bigint<br>Integer | bigint<br>Integer | string<br>Date (unparsed) | string<br>Text | string<br>Text | bigint<br>Integer | string<br>Text | double<br>Decimal | bigint<br>Integer |
| 59760 | 1 | 2017-05-08 14:19:45 | C_ID_e0fd181776 | M_ID_1eb5b5e073 | 417 | B | 34.4 | 0 |
| 59761 | 1 | 2017-05-08 14:22:04 | C_ID_e7676db3b8 | M_ID_7ca715839e | 771 | D | 99.3 | 1 |
| 59762 | 1 | 2017-05-08 14:24:45 | C_ID_a15f532627 | M_ID_4885bb4884 | 518 | D | 359.25 | 0 |
| 59763 | 1 | 2017-05-08 14:25:12 | C_ID_3092725553 | M_ID_b8a0b209fe | 206 | B | 60.08 | 1 |
| 59764 | 1 | 2017-05-08 14:25:47 | C_ID_5c32762ada | M_ID_e1b2d6bf5c | 373 | B | 241.27 | 0 |
| 59765 | 1 | 2017-05-08 14:26:27 | C_ID_45e7b34f81 | M_ID_d5c94df60e | 434 | B | 418.73 | 0 |
| 59766 | 1 | 2017-05-08 14:27:47 | C_ID_17226c9d33 | M_ID_6f939b9049 | 278 | C | 51.73 | 0 |
| 59767 | 1 | 2017-05-08 14:28:47 | C_ID_9dd08a4854 | M_ID_b86f065dde | 783 | C | 90.41 | 0 |

Each row is a transaction. The ‘authorized_flag’ shows whether it was approved or not.

# Column attributes




**Column name**

**Schema** - the storage type

ex. string, int, double, array

**Meaning** - the "human readable" meaning

ex. text, integer, IP address, gender

We can see information about each column and its contents

# Agenda

Discovery
Hands-on Exercise

1. Background
2. DSS Login and Set up Account
3. Creating Projects and Connecting to Data
4. **Stacking Datasets**
5. Joins
6. Computation Engines
7. Data Cleaning [1]
8. Charts [1]
9. Splitting Data
10. Machine Learning

Green Track

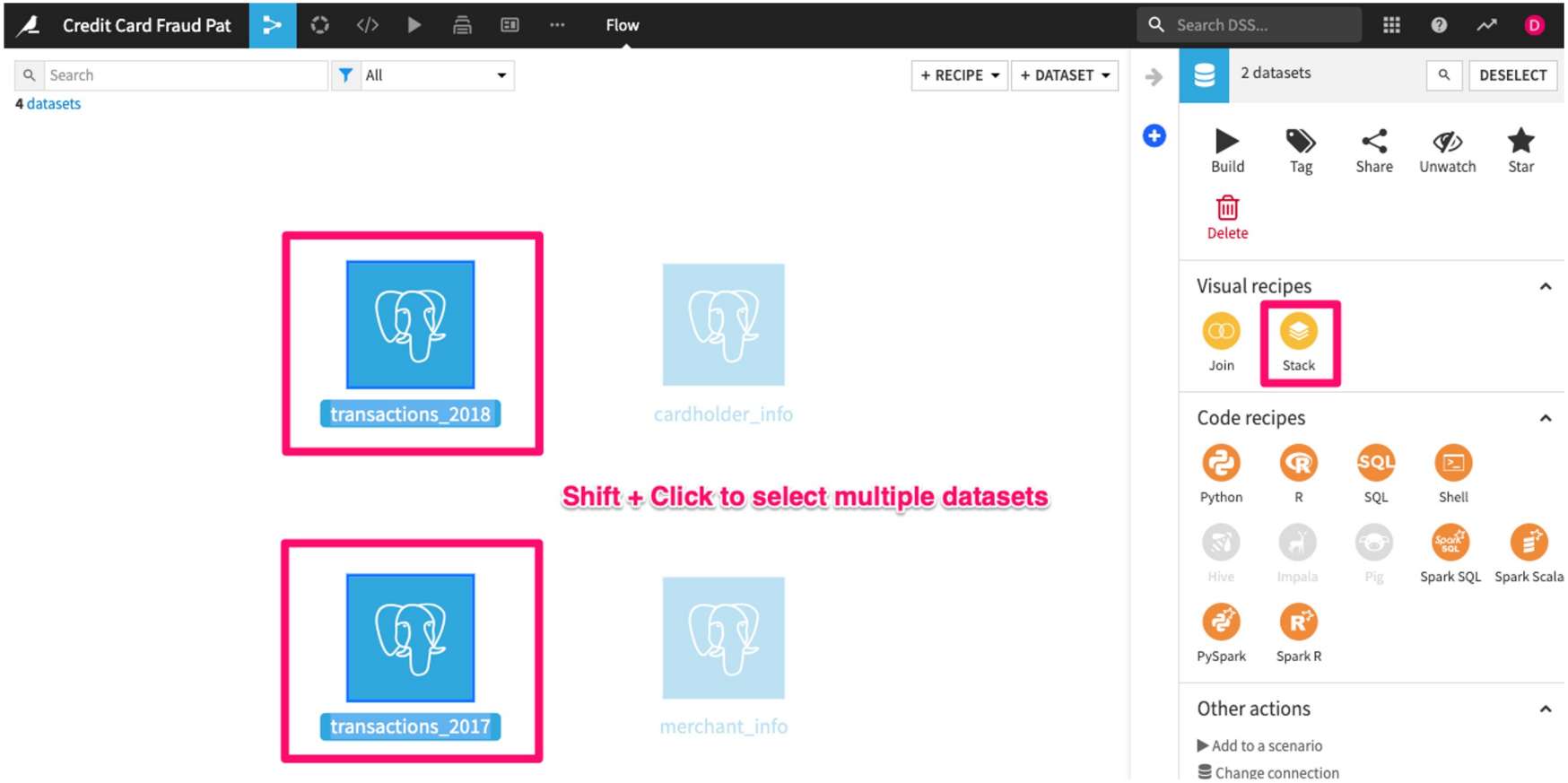
Purple Track

11. Filtering Data
12. Data Cleaning [2]
13. Group by Aggregations
14. Charts [2]

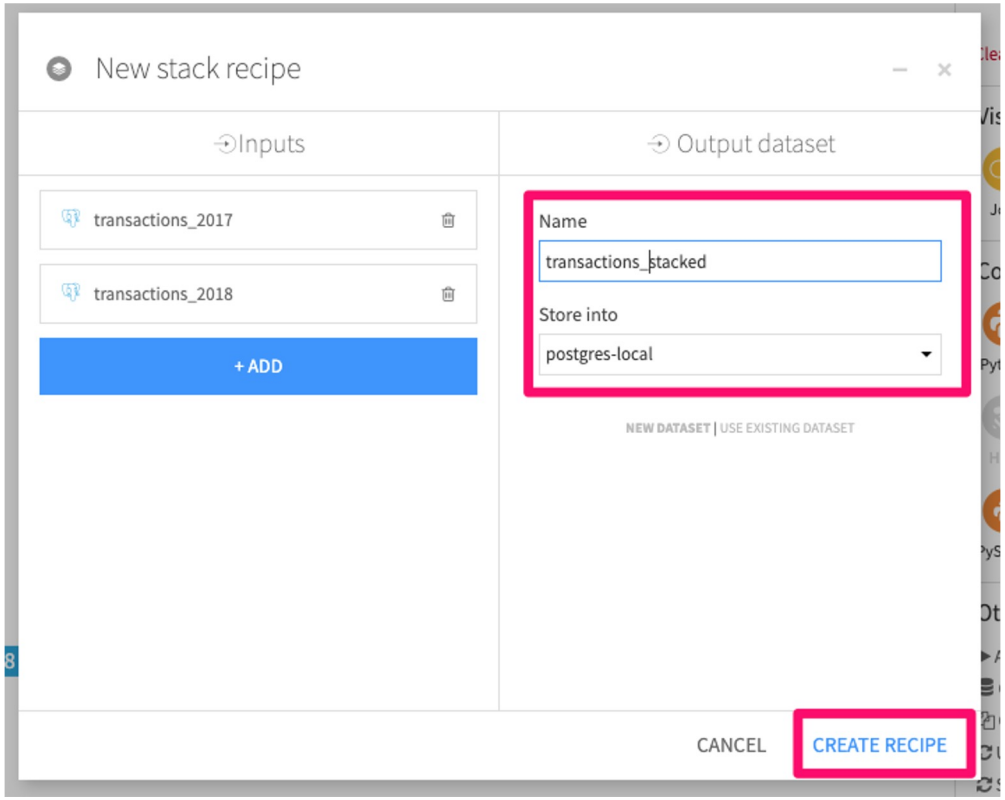
# Stack

- Merge two or more datasets into one
- Equivalent to a “union all” SQL statement
- Good for datasets with substantially the same columns/schema

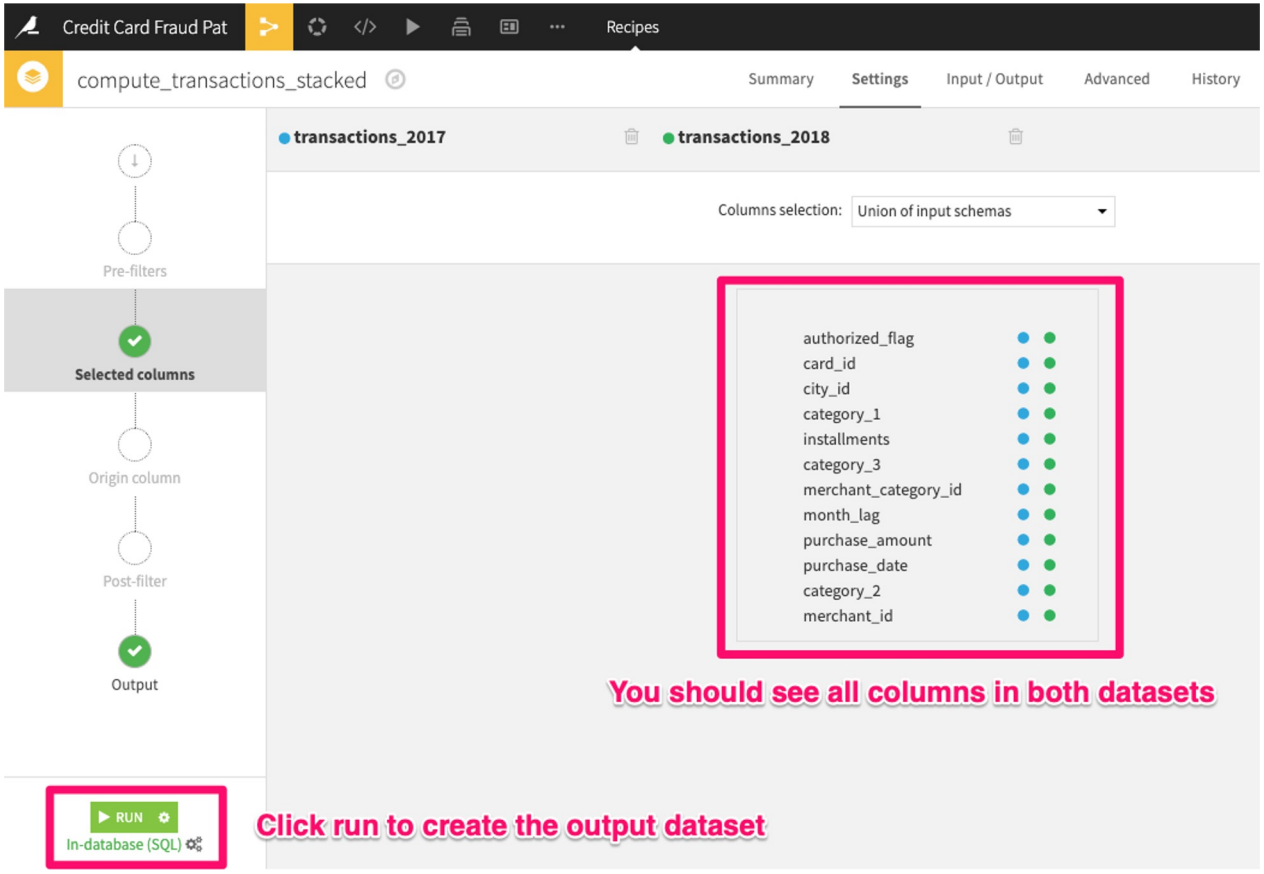
# Stack the transactions datasets together



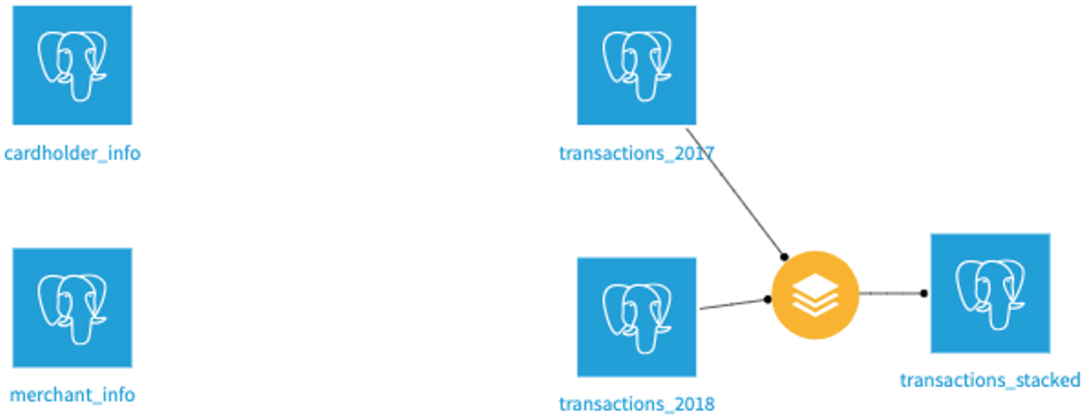
# Rename the output dataset “transactions_stacked”



# Stack


Credit Card Fraud Pat
Recipes
compute_transactions_stacked
Summary
Settings
Input / Output
Advanced
History
transactions_2017
transactions_2018
Columns selection:
Union of input schemas
Pre-filters
Selected columns
Origin column
Post-filter
Output
authorized_flag
card_id
city_id
category_1
installments
category_3
merchant_category_id
month_lag
purchase_amount
purchase_date
category_2
merchant_id
You should see all columns in both datasets
RUN
In-database (SQL)
Click run to create the output dataset

# Your flow should look like this



# Agenda

Discovery
Hands-on Exercise

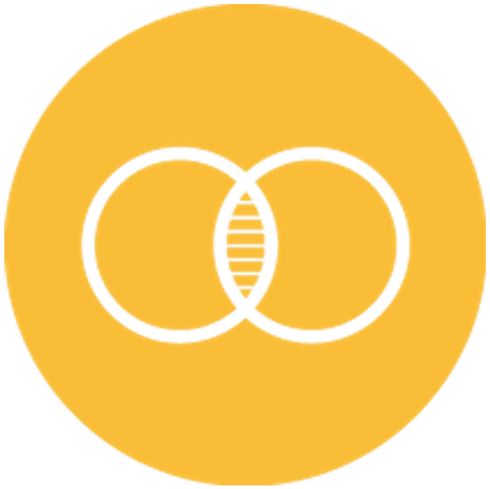
1. Background
2. DSS Login and Set up Account
3. Creating Projects and Connecting to Data
4. Stacking Datasets
5. Joins
6. Computation Engines
7. Data Cleaning [1]
8. Charts [1]
9. Splitting Data
10. Machine Learning

Green Track

Purple Track

11. Filtering Data
12. Data Cleaning [2]
13. Group by Aggregations
14. Charts [2]

# Join



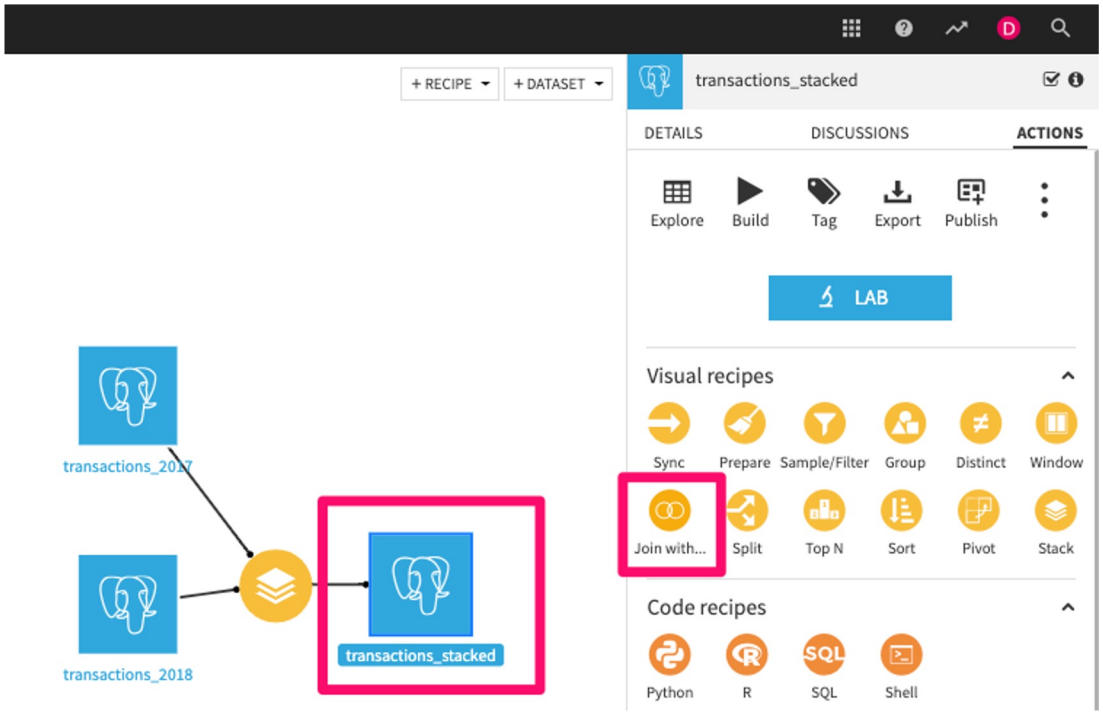
- Join two or more datasets
- Perform inner joins, left outer joins, full outer joins, right joins, etc.
- Define join conditions using the UI or SQL

# Join the stacked transactions dataset with the other two datasets

1. LEFT JOIN transactions_stacked dataset with cardholder_info dataset ON ‘card_id’ = ‘internal_card_mapping’
2. LEFT JOIN transactions dataset with merchant_info dataset ON ‘merchant_id’ and ‘merchant_category_id’
3. ADD PREFIXES “cardholder” and “merchant” to columns from the two datasets respectively

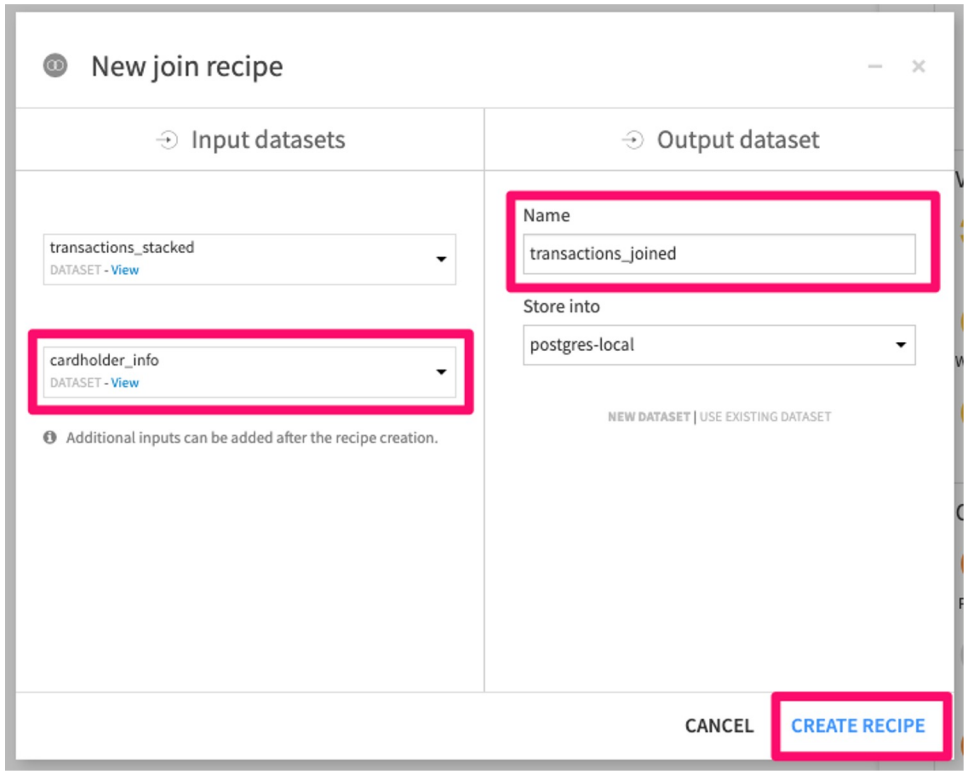
Do this all in ONE Join recipe

# Join the stacked transactions dataset with the other two datasets



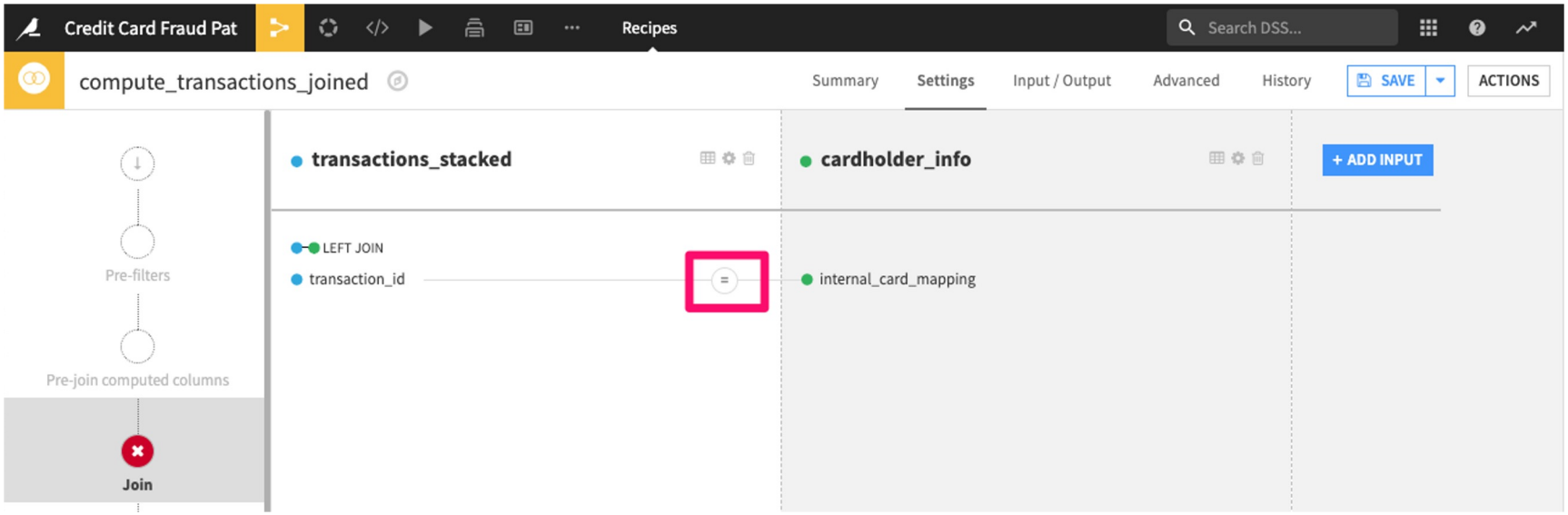

- Click on the transactions_stacked dataset
- Choose the Join recipe

# Join the stacked transactions dataset with the other two datasets




- Select the cardholder_info dataset from the dropdown
- Name the output dataset transactions_joined

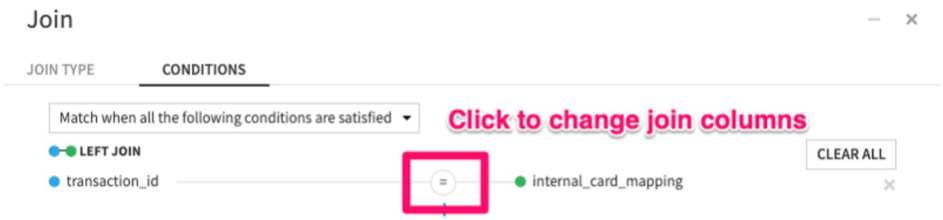
# LEFT JOIN transactions_stacked with cardholder_info




Sometimes DSS guesses the wrong desired join columns

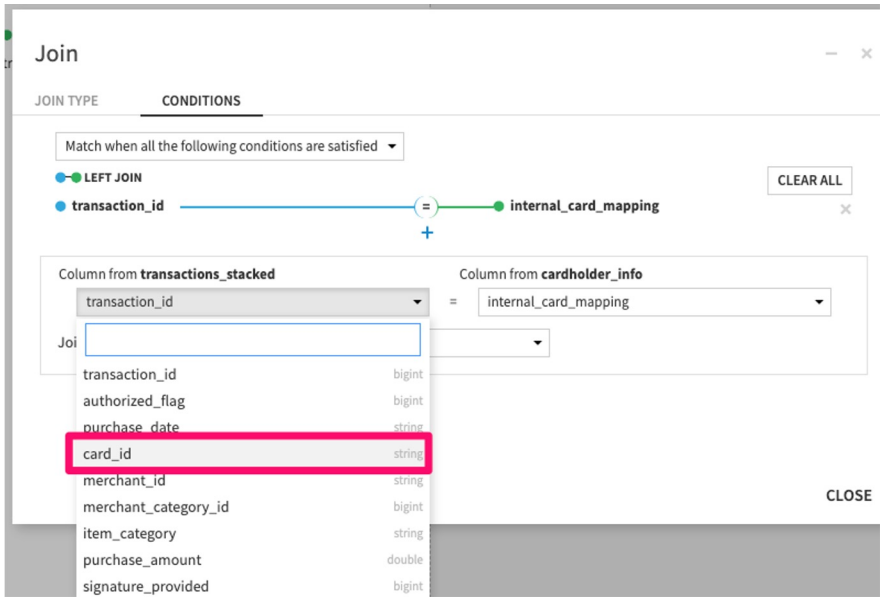
Click the equals sign to change the join

# LEFT JOIN transactions_stacked with cardholder_info

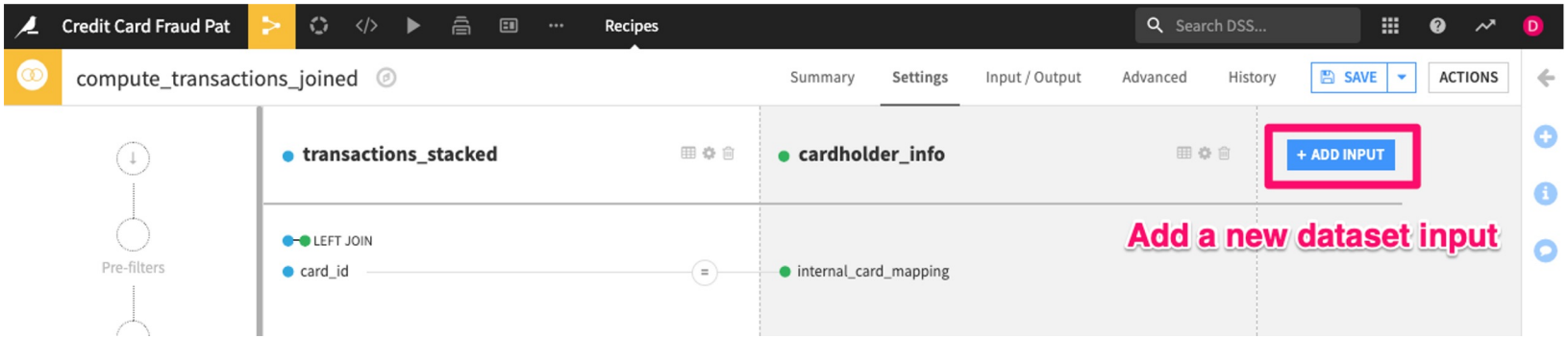



Change the columns to join on

'card_id' = 'internal_card_mapping'

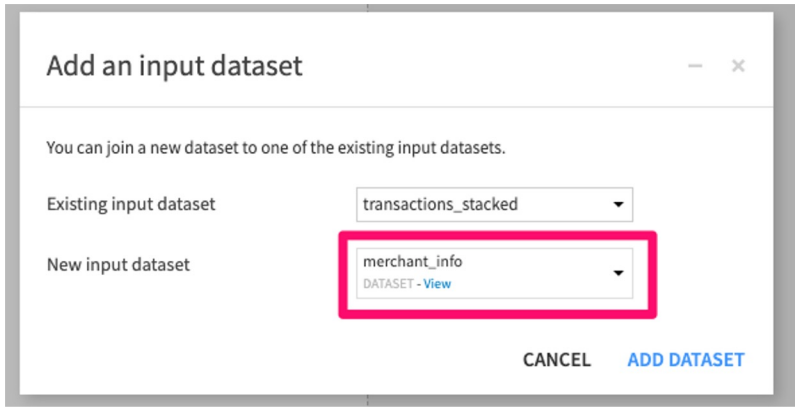


# LEFT JOIN transactions_stacked with merchant_info

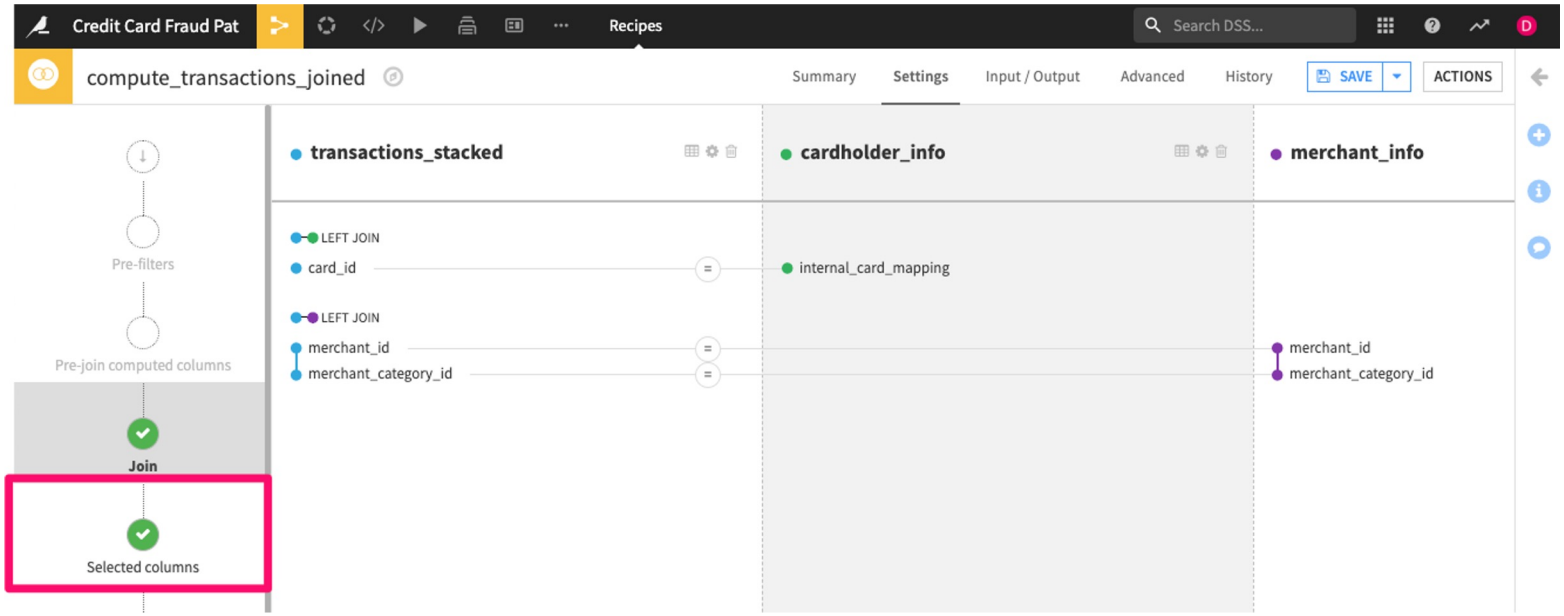



Add a new dataset input

Select the merchant_info dataset from the dropdown



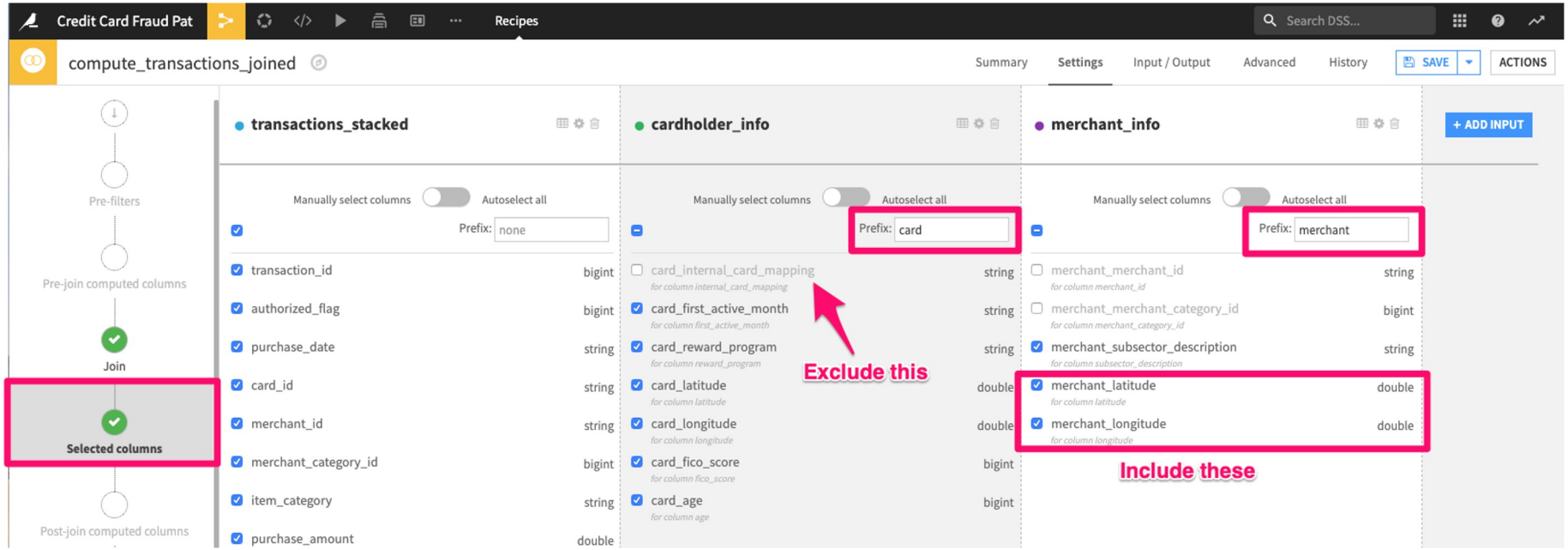
# LEFT JOIN transactions_stacked with merchant_info




This looks good!

Now let's look at the Selected Columns tab

# Add prefixes to columns from joined datasets

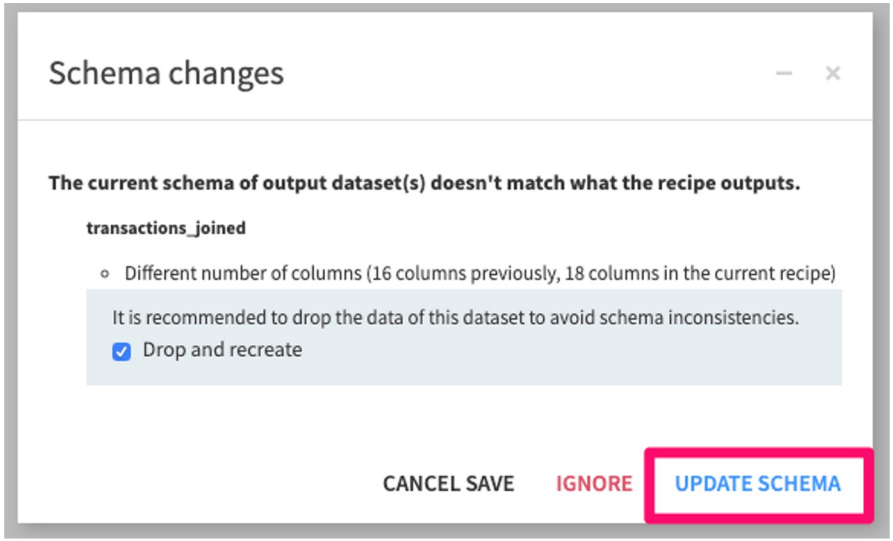



- Deselect ‘card_internal_card_mapping’
- Select ‘merchant_latitude’ and ‘merchant_longitude’

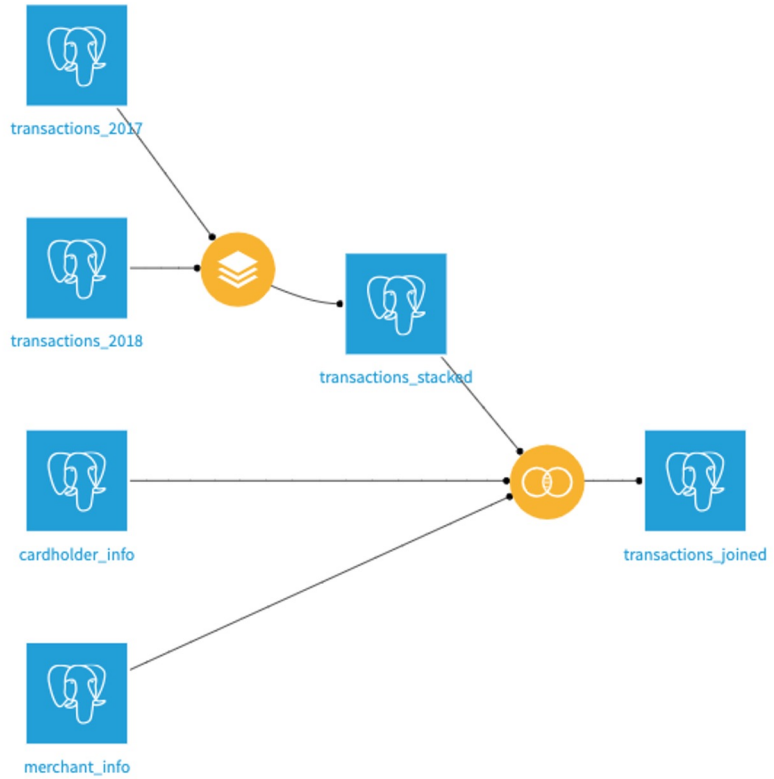
Any ideas why merchant latitude/longitude cols weren’t included initially?

# Run your join


Don’t be afraid of schema updates! (Click the blue button)



# Flow check


transactions_2017
transactions_2018
transactions_stacked
cardholder_info
transactions_joined
merchant_info

# Agenda

Discovery
Hands-on Exercise

1. Background
2. DSS Login and Set up Account
3. Creating Projects and Connecting to Data
4. Stacking Datasets
5. Joins
6. **Computation Engines**
7. Data Cleaning [1]
8. Charts [1]
9. Splitting Data
10. Machine Learning

Green Track

Purple Track

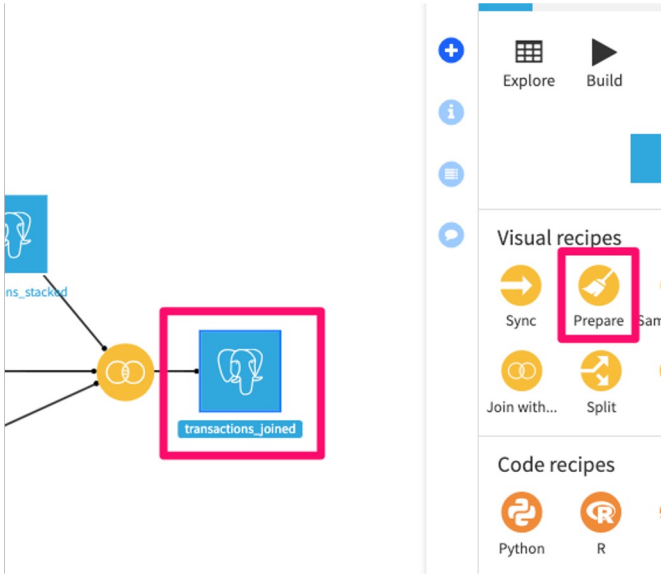
11. Filtering Data
12. Data Cleaning [2]
13. Group by Aggregations
14. Charts [2]

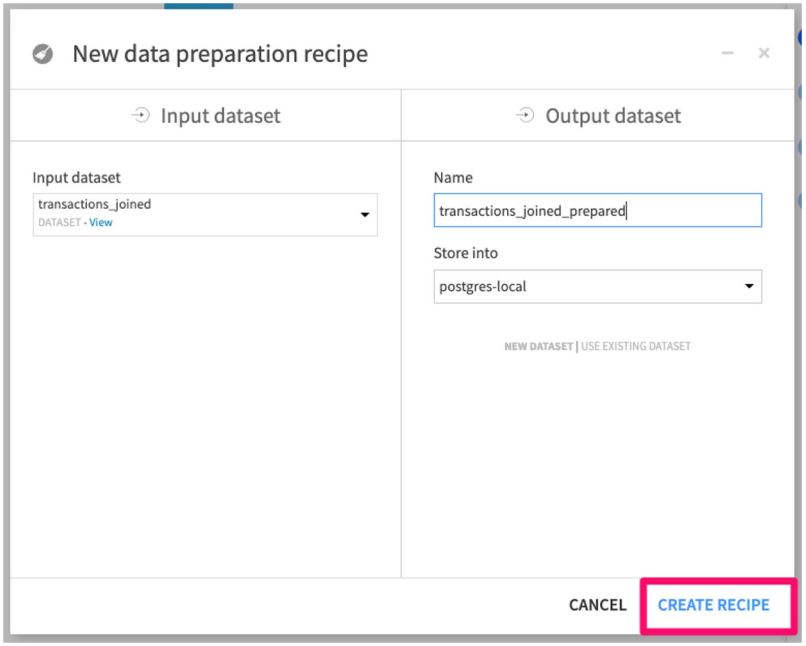
# Prepare

~90 different processors, including

- Find and replace
- Date manipulations
- Mathematical operations on columns
- Splitting columns
- Extracting parts of columns
- Text parsing and NLP
- Geographic transformations

# Create a Prepare recipe from the transactions_joined dataset

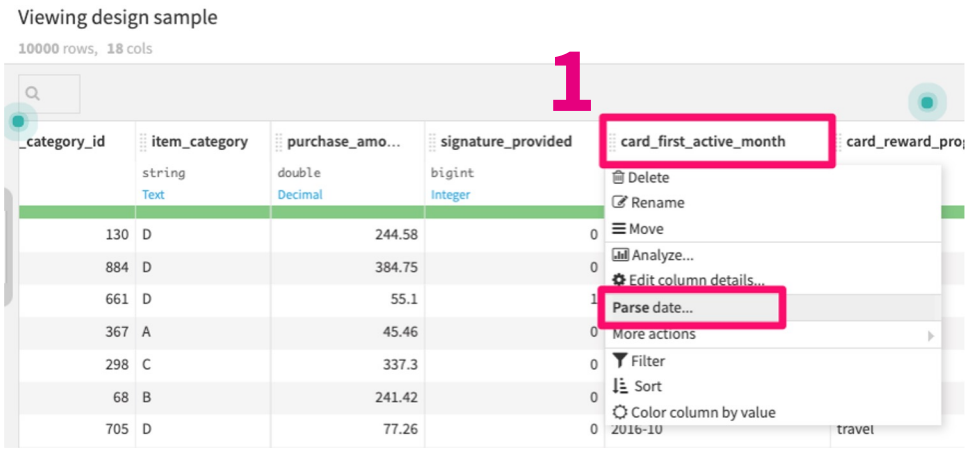


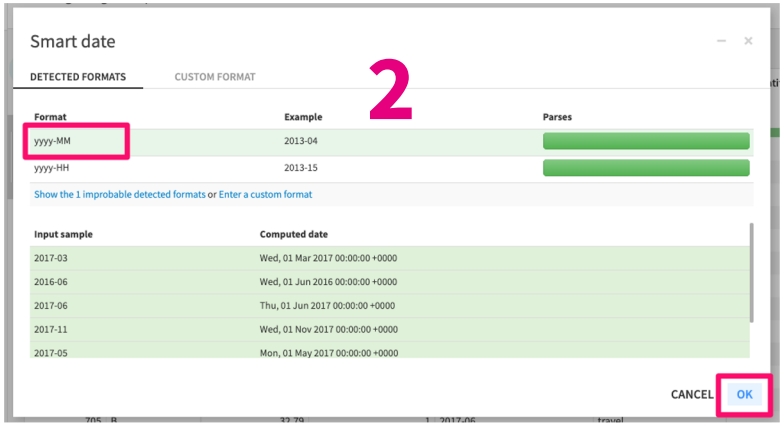



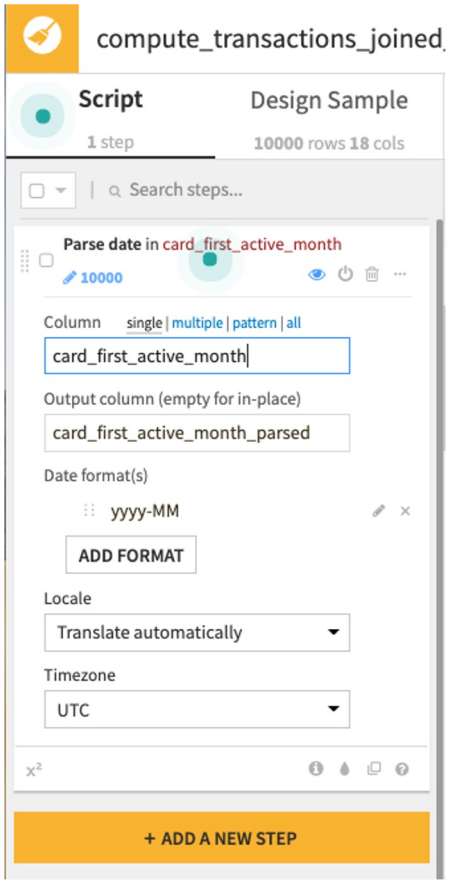
## Round 1: Create time-based features

1. Parse dates: card_first_active_month and purchase_date
2. Extract date components from purchase_date_parsed
    - year, month, day, dow, hour
3. Create a purchase_weekend (sat/sun) column using a DSS formula
4. Create a days_active column: the time from the cardholder's first_active_month until purchase_date

# Parse card_first_active_month



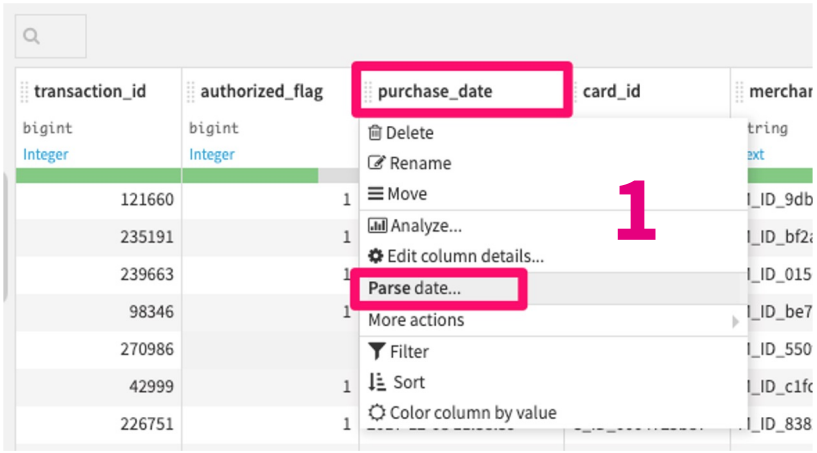



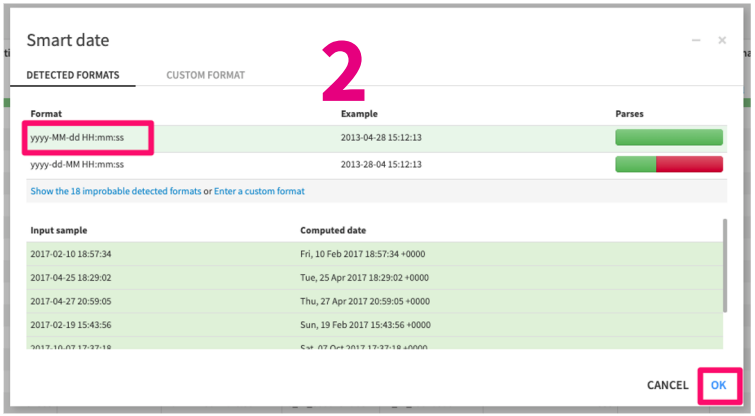



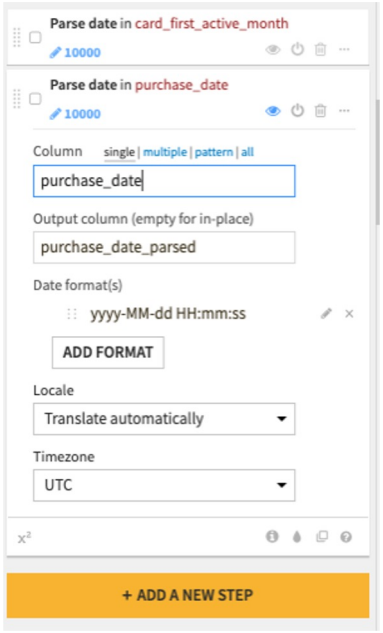

3

We want yyyy-MM format

# Parse purchase_date

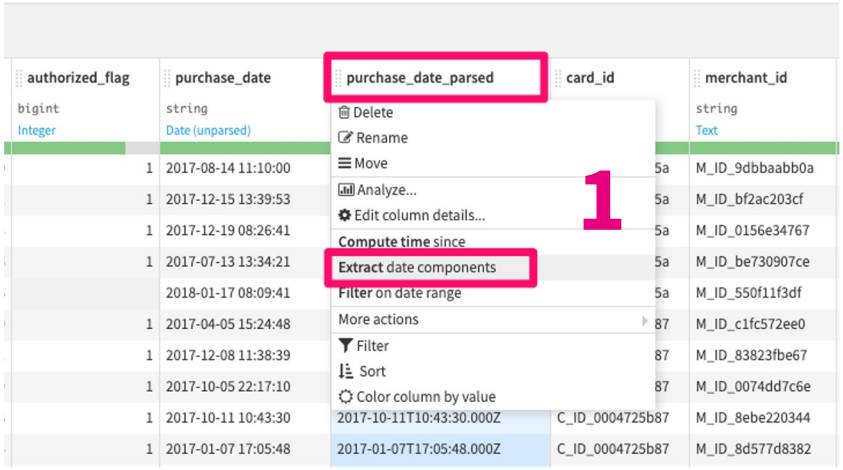


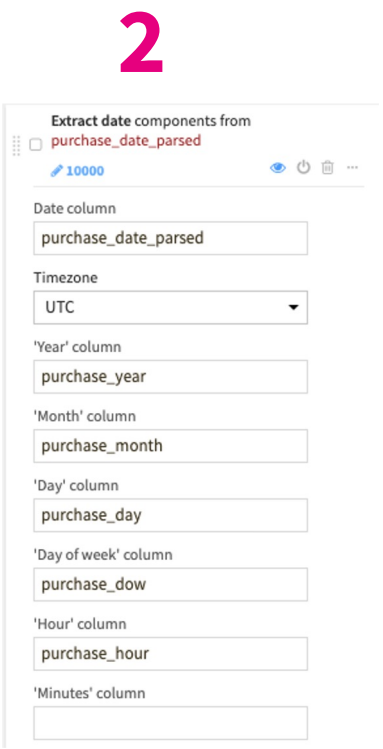




**Note**: if you wanted to parse in-place, you could delete the output column name to do that

# Extract date components from purchase_date_parsed



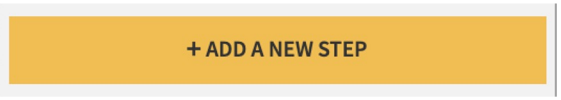

Clean up column names:
for example…
purchase_date_parsed_year → purchase_year

Add:
purchase_dow
purchase_hour

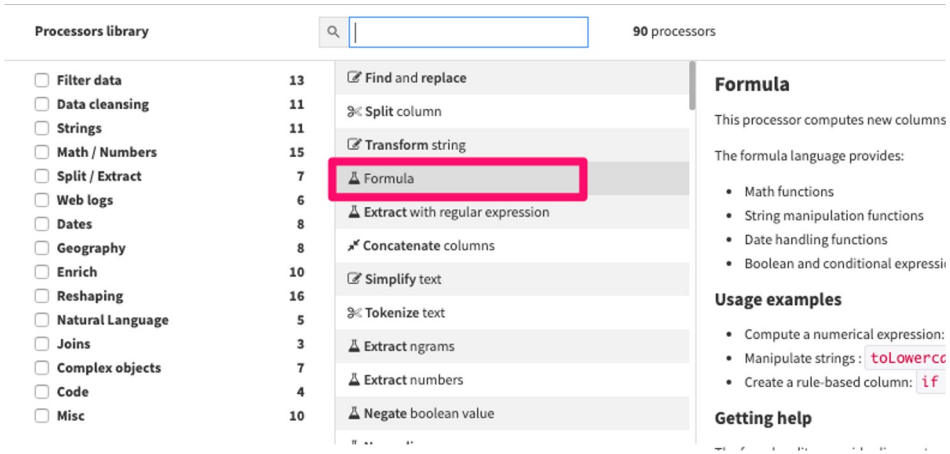
# Create binary purchase_weekend column

Use a DSS Formula!

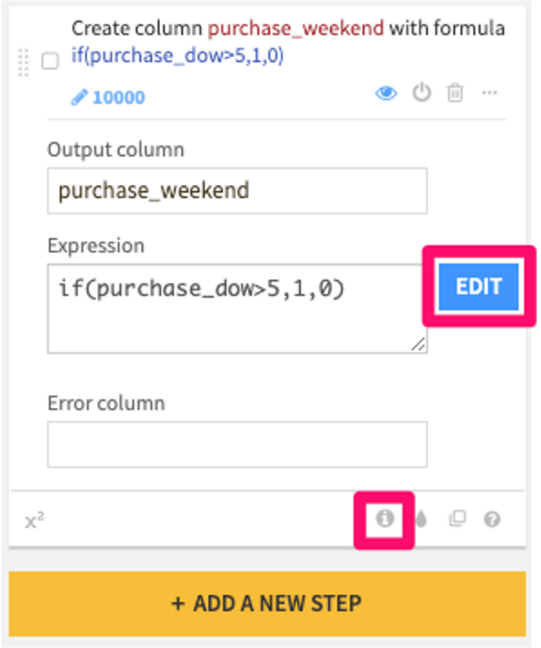
We are defining weekends
as Friday/Saturday




Let's add this one using the
+ ADD A NEW STEP button

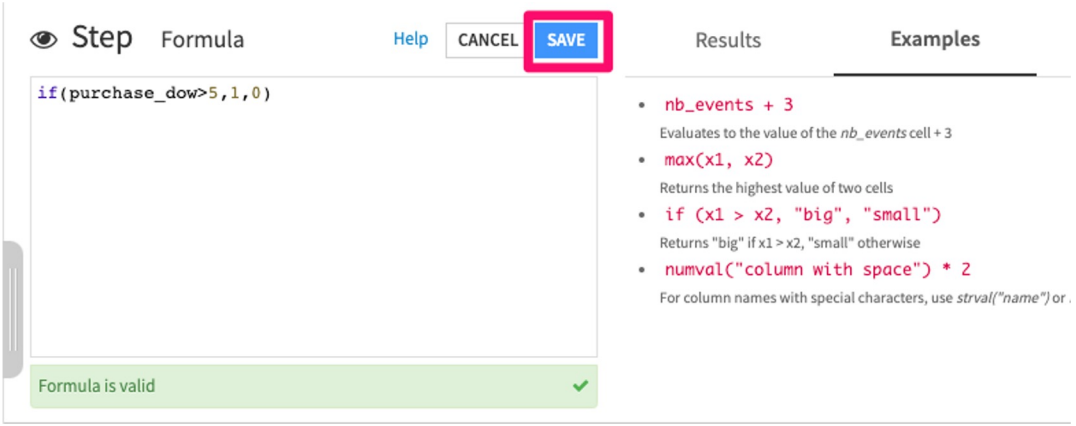


# Create binary purchase_weekend column

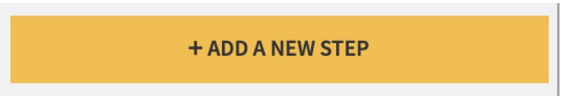



You can add comments to document potentially complex formulas!

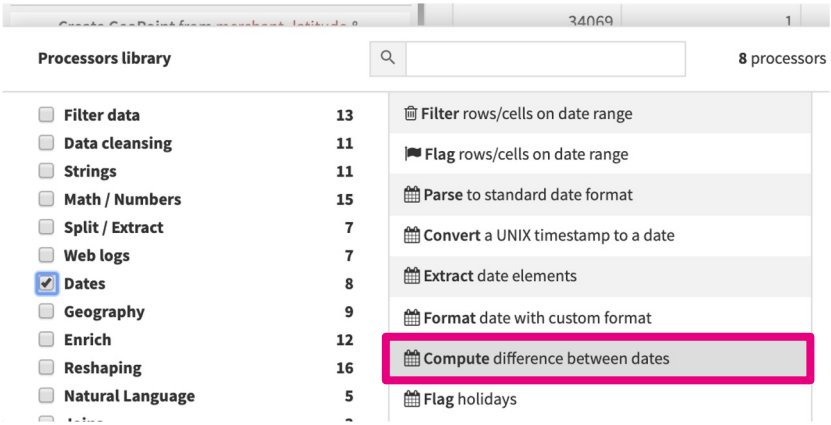
Click "EDIT" to pull up examples and reference

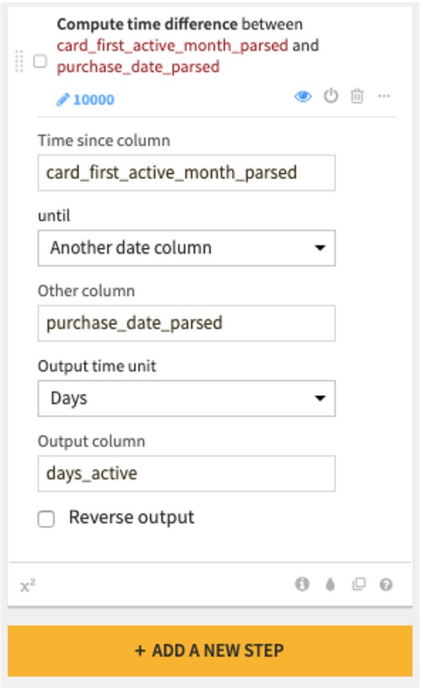


# Compute the time from first active month to purchase date



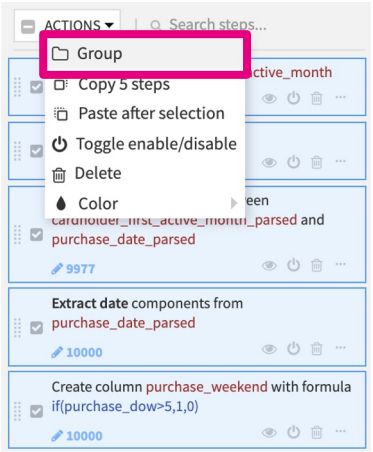

Let’s add this one using the + ADD A NEW STEP button

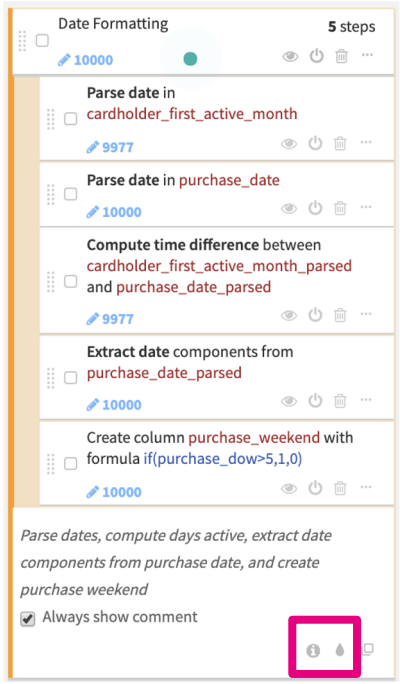



Name the output column: days_active

# You can Group your date cleaning steps




Call your group “Date Formatting”

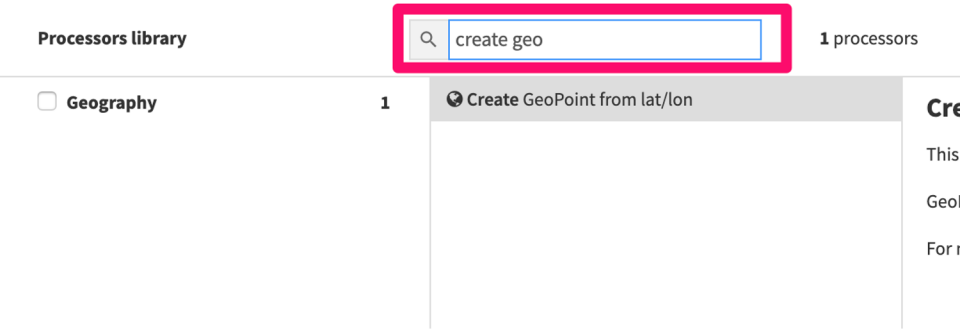
This is much easier to read!

**You can also modify colors or add descriptions !**

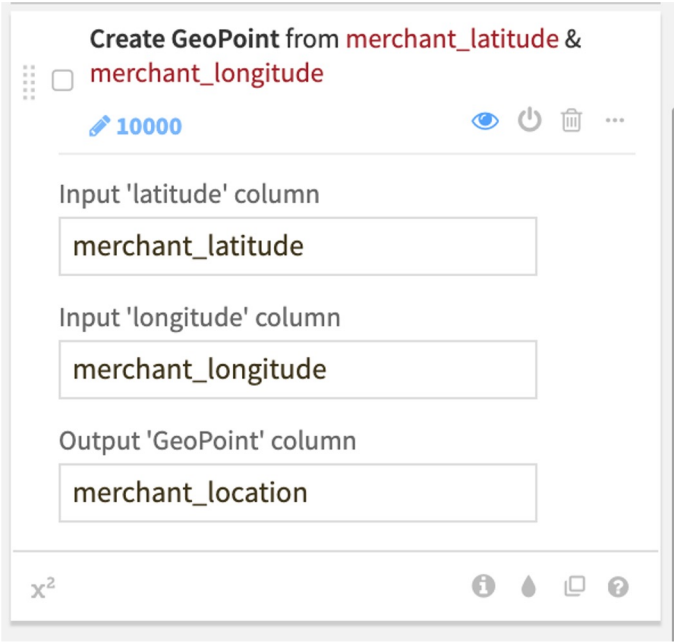
## Round 2: Create geolocation features

1. Create geopoint columns
    - merchant_location using the merchant latitude & longitude
    - cardholder_location using the cardholder latitude & longitude
2. Reverse geocode
    - merchant_location to get the merchant_state
    - cardholder_location to get the cardholder_state

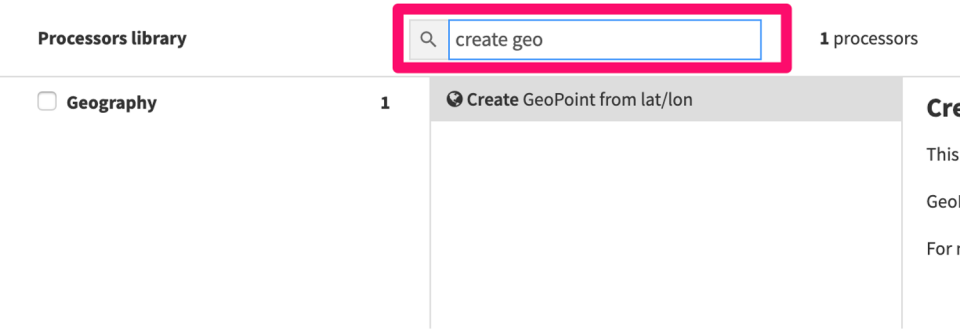
# Create merchant geopoint column

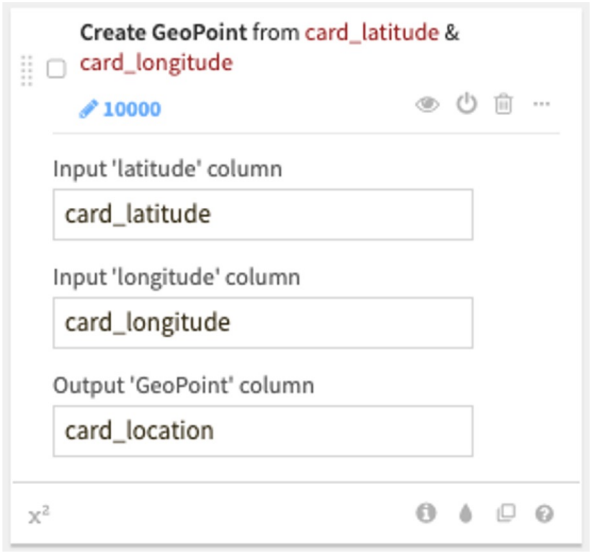



Use the processor search



# Create cardholder geopoint column



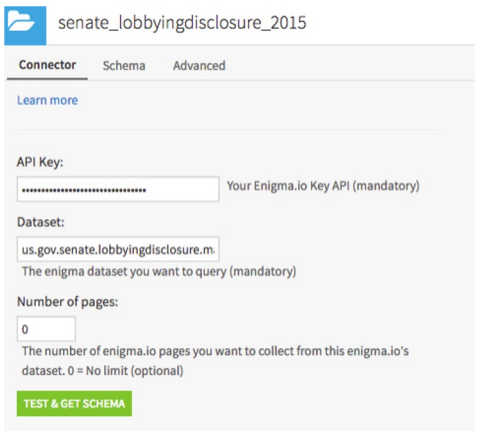




Do the same thing for the cardholder locations!

For step 2, we will be using the reverse geocoding plugin…

# Plugins

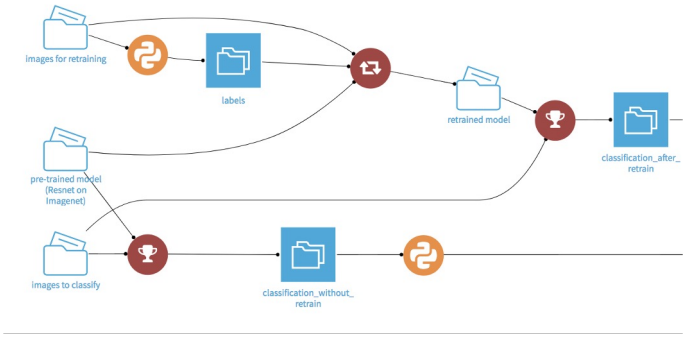
## Many Types




In DSS plugins, the user can develop :

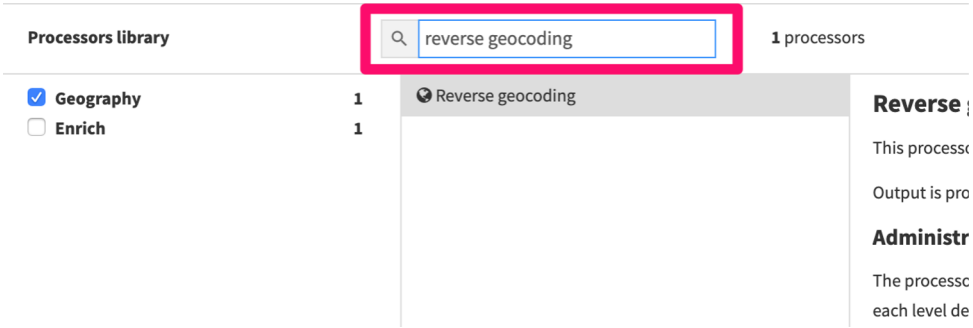
- **Custom Recipes**
  - A basic user interface allowing the users to interact with variables of the underlying code: R, Python, SQL in Python
  - Example : Query a rest API with some settings; apply a statistical function to a field of dataset
- **Custom Dataset**
  - Example : query a salesforce table.
- **Much More**
  - New Preparation Step Processor
  - Custom Scenario Step
  - Dataset Exporter
  - Web App/R Markdown Template

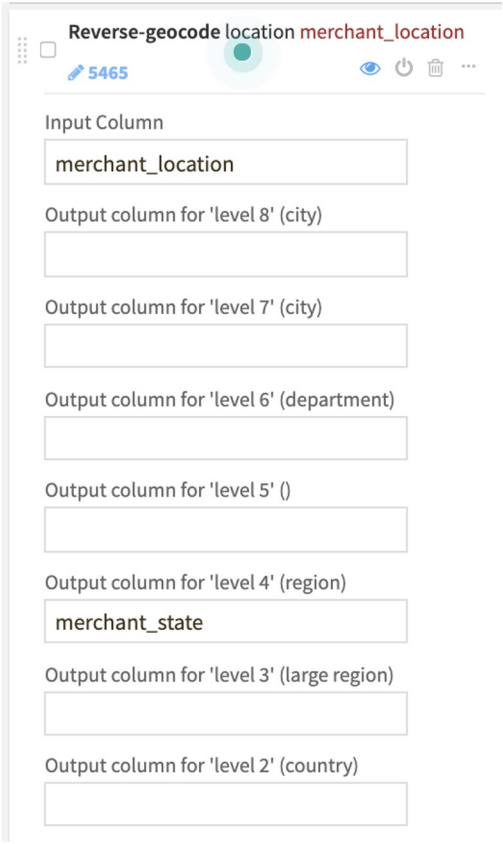
Plugins are then available as a DSS extension



# Reverse geocode the merchant location

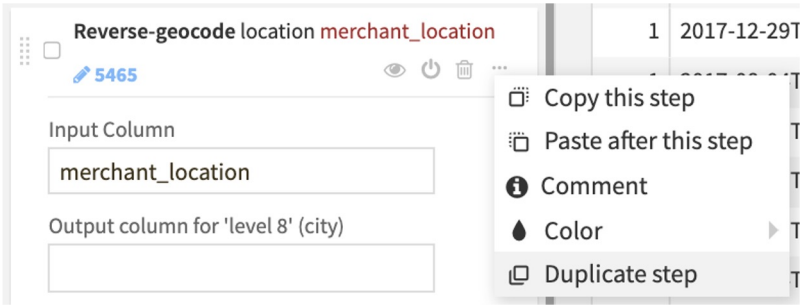
Use the processor search



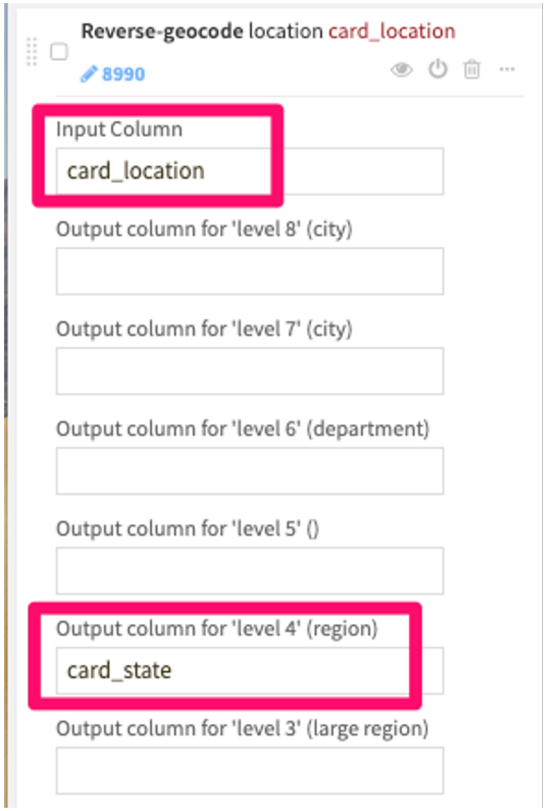




We can choose what to extract.
In “region”, enter “merchant_state”

# Reverse geocode the cardholder location

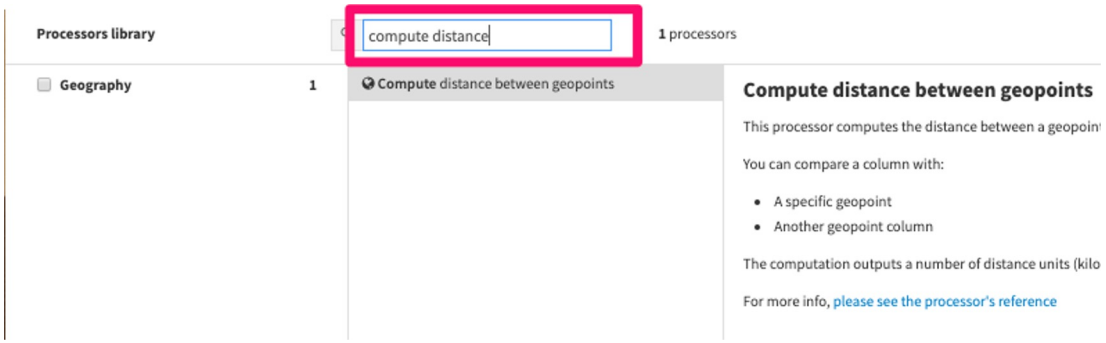


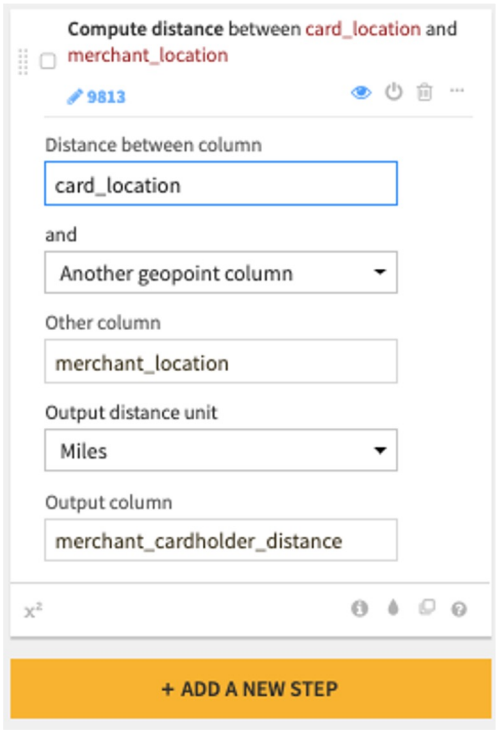

Feeling lazy? Duplicate and modify!



# Compute the distance between cardholders and merchants

Use the processor search

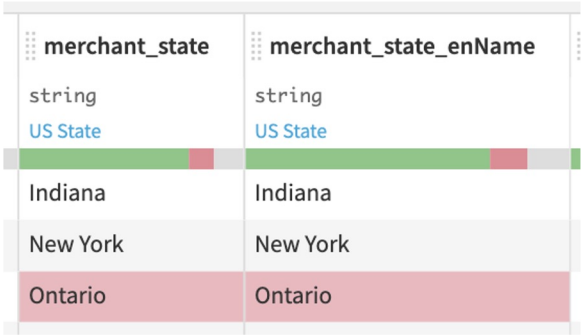


# Note the “invalid” data

DSS has two “data types” listed:

1) Storage Type: indicates how dataset backend should store data
2) Meaning: rich semantic type that can be more specific and is automatically detected

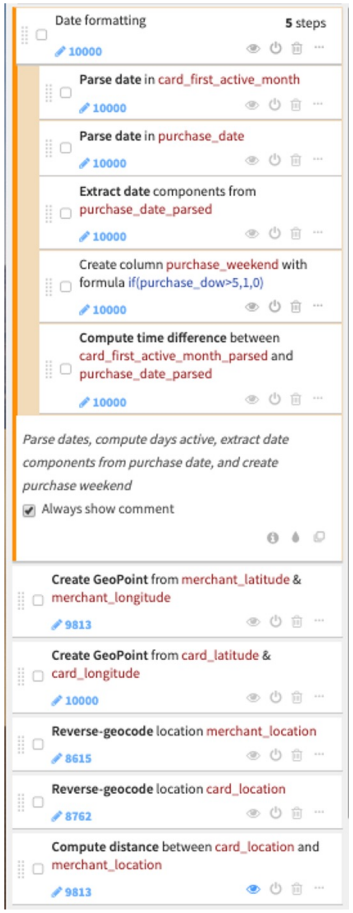


| merchant_state | merchant_state_enName |
|---|---|
| string | string |
| US State | US State |
| Indiana | Indiana |
| New York | New York |
| Ontario | Ontario |

The red refers to invalid meanings. These will not necessarily cause errors, they are more to help guide you.

Including non-US data is fine for now, so we can leave the “invalid” data in.

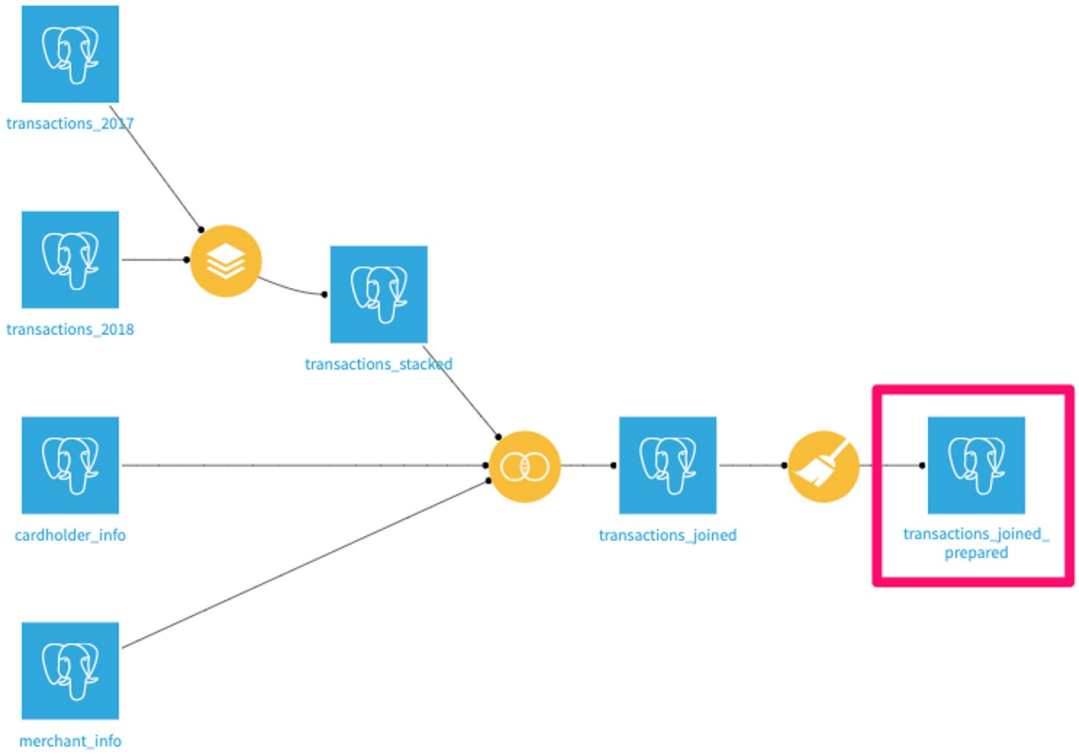
# Double check that you've added all steps



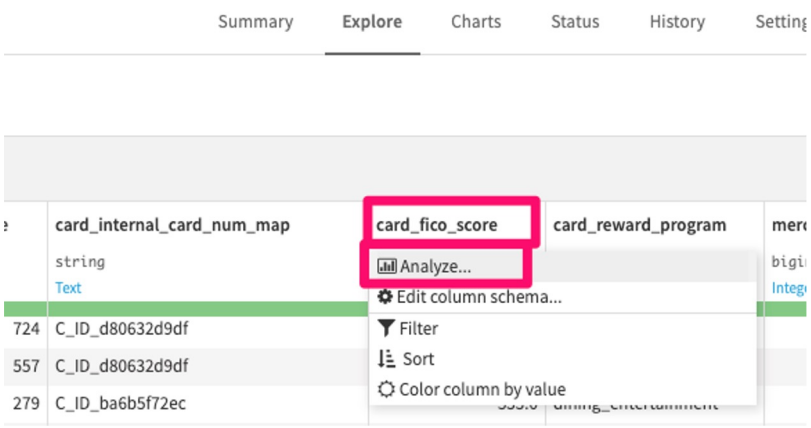


...and then click "RUN"

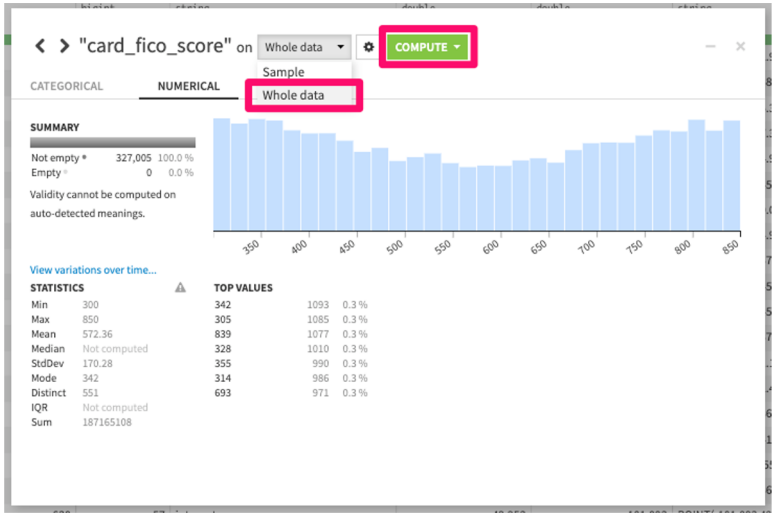
# Analyze the transactions_joined_prepared dataset



# Click on a column name and ‘Analyze’ to see more




- Distributions
- Summary stats
- Value counts



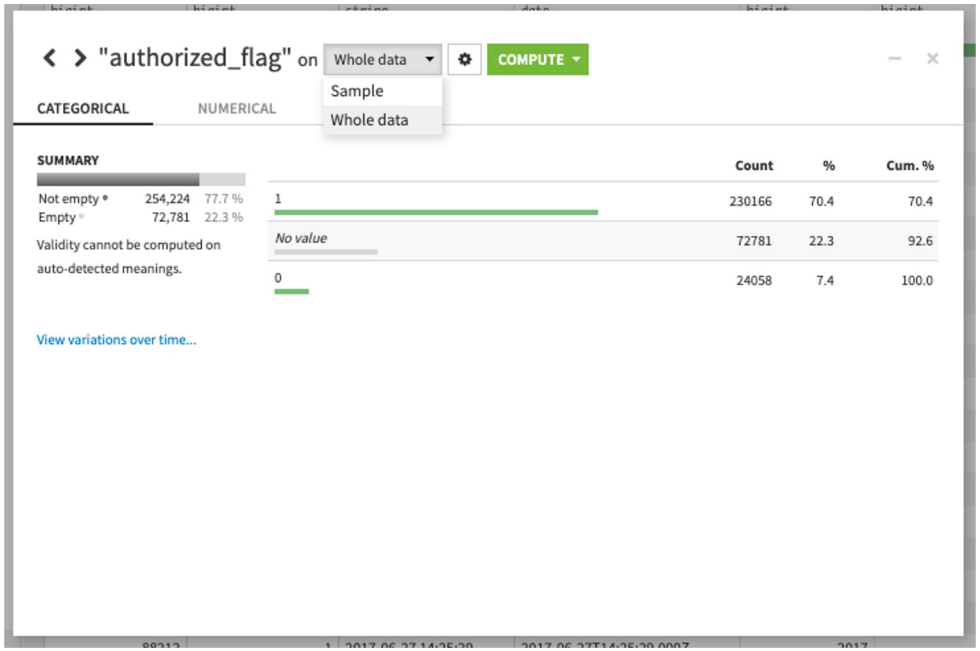

Stats can be computed on the sample or whole dataset

# Questions

1. What is the highest individual purchase amount?

1. What percentage of transactions are unauthorized (fraudulent)?

# When analyzing a column, you can run stats on the entire dataset



# Questions

1. What is the highest individual purchase amount?
   $48,147
1. What percentage of transactions are unauthorized?
   7.4%

# Agenda

Discovery
Hands-on Exercise

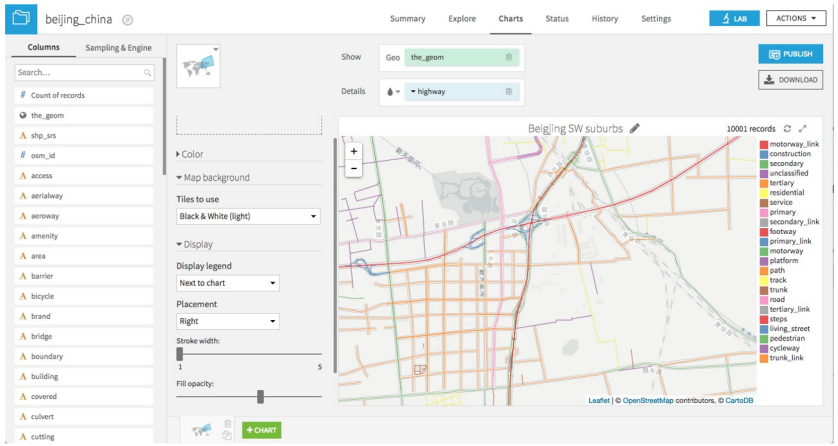
1. Background
2. DSS Login and Set up Account
3. Creating Projects and Connecting to Data
4. Stacking Datasets
5. Joins
6. Computation Engines
7. Data Cleaning [1]
8. Charts [1]
9. Splitting Data
10. Machine Learning

Green Track

Purple Track

11. Filtering Data
12. Data Cleaning [2]
13. Group by Aggregations
14. Charts [2]


Insight Sum of revenue_loss by cluster_labels
Summary
View
Edit
BACK TO DASHBOARD
GO TO SOURCE DATASET
ACTIONS
Display
Display legend
Next to chart
Placement
Right
Display value in chart
Color
Animation
Subcharts
intl_Plan
Chart height
50
229
300
Common X Axis
Display tooltip on each subchart
Run: In DSS
Sum of revenue_loss by cluster_labels
1809 records
day_callers
eve_night_callers
high_spenders
local_callers
low_spenders
vmailers
no
yes
5000
4000
3000
2000
1000


beijing_china
Summary
Explore
Charts
Status
History
Settings
LAB
ACTIONS
Columns
Sampling & Engine
Search...
Count of records
the_geom
shp_srs
osm_id
access
aerialway
aeroway
amenity
area
barrier
bicycle
brand
bridge
boundary
building
covered
culvert
cutting
Show
Geo
the_geom
Details
highway
PUBLISH
DOWNLOAD
Beijing SW suburbs
10001 records
Color
Map background
Tiles to use
Black & White (light)
Display
Display legend
Next to chart
Placement
Right
Stroke width:
1
5
Fill opacity:
motorway_link
construction
secondary
unclassified
tertiary
residential
service
primary
secondary_link
footway
primary_link
motorway
platform
path
track
trunk
road
tertiary_link
steps
living_street
pedestrian
cycleway
trunk_link
Leaflet | © OpenStreetMap contributors, © CartoDB
CHART

# Build a chart on the transactions_stacked_prepared dataset

Task:

- Map the fraud rate by US state
- Which states have the highest rates?

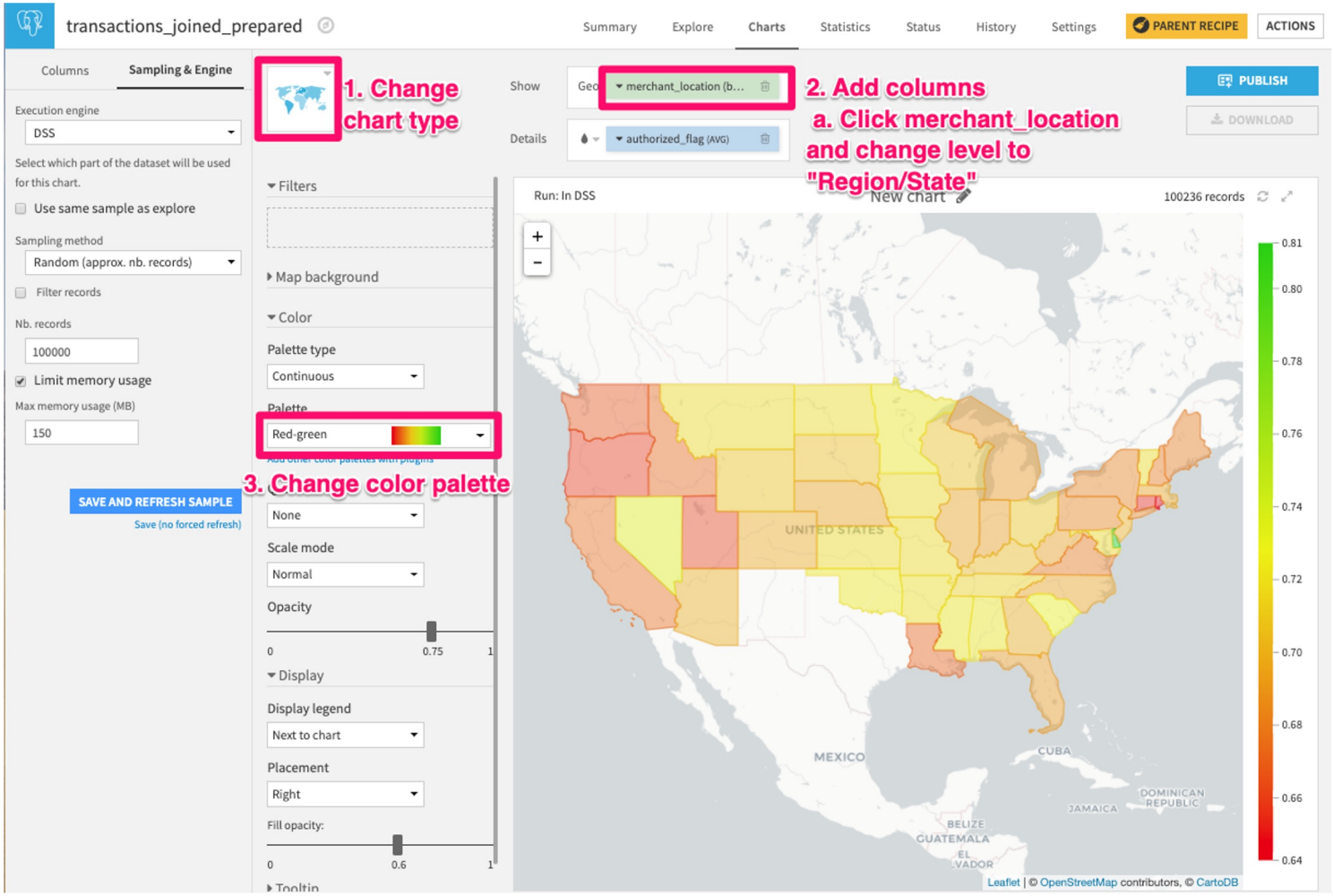
Details:

- Filled administrative map
- Geo - merchant_location
- Fill - authorized_flag (AVG)

Note:

- Change the sample to 100,000 random (left hand menu)

# Maps


transactions_joined_prepared
Summary
Explore
Charts
Statistics
Status
History
Settings
PARENT RECIPE
ACTIONS
Columns
Sampling & Engine
Execution engine
DSS
Select which part of the dataset will be used for this chart.
Use same sample as explore
Sampling method
Random (approx. nb. records)
Filter records
Nb. records
100000
Limit memory usage
Max memory usage (MB)
150
SAVE AND REFRESH SAMPLE
Save (no forced refresh)
1. Change chart type
Show
merchant_location (b...
Details
authorized_flag (AVG)
2. Add columns
a. Click merchant_location and change level to "Region/State"
PUBLISH
DOWNLOAD
Filters
Map background
Color
Palette type
Continuous
Palette
Red-green
3. Change color palette
None
Scale mode
Normal
Opacity
0
0.75
1
Display
Display legend
Next to chart
Placement
Right
Fill opacity:
0
0.6
1
Tooltip
Run: In DSS
New chart
100236 records
UNITED STATES
MEXICO
CUBA
JAMAICA
DOMINICAN REPUBLIC
BELIZE
GUATEMALA
EL VADOR
0.81
0.80
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.64
Leaflet | © OpenStreetMap contributors, © CartoDB

# Scatter plots

Task:

- Plot the purchase_amount by FICO score - show which transactions are fraudulent
- Does it look like people with higher FICO scores have more fraudulent transactions?
- What about high vs. low transaction amounts?

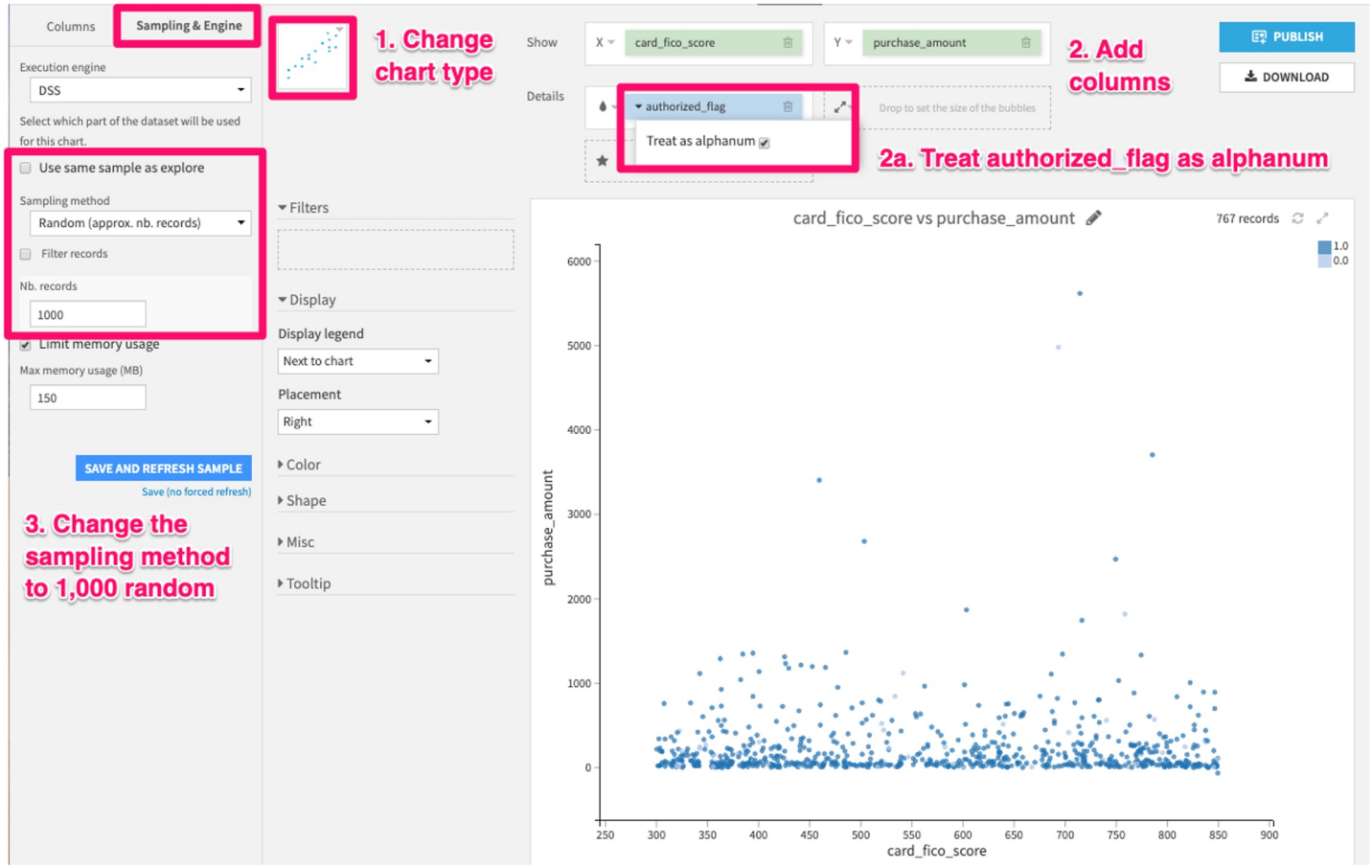
Details:

- Scatter plot
- X-axis - card_fico_score
- Y-axis - purchase_amount
- Color - authorized_flag

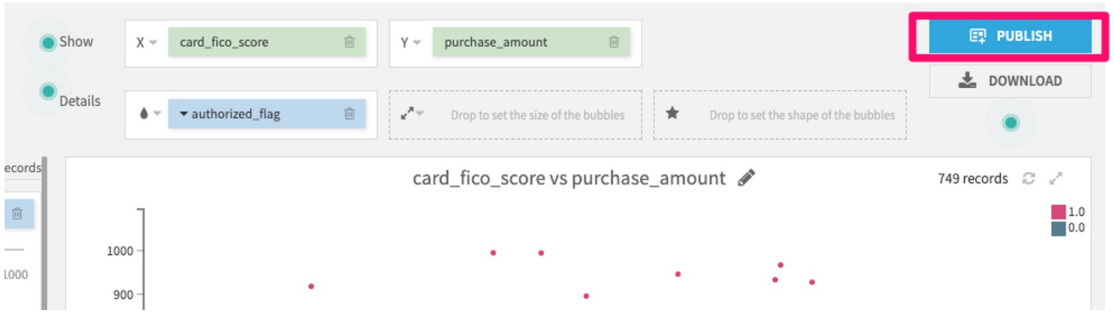
Note:

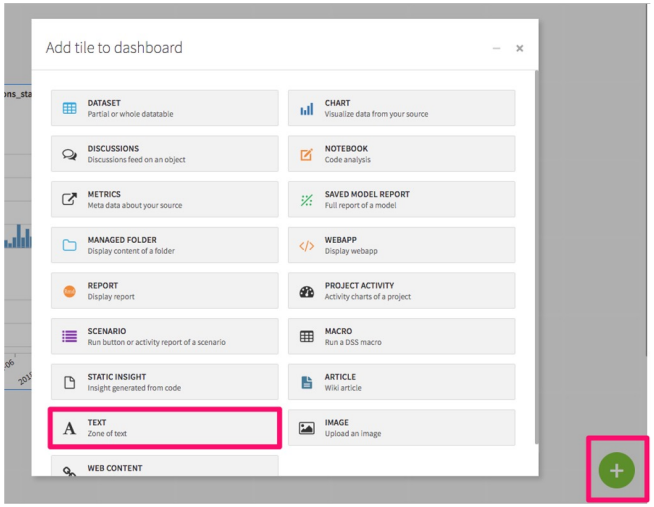
- Change the sample to 1,000 random (left hand menu)
- Filter for only purchases less than $1000

# Scatter plots



# Publish both of your charts to a dashboard

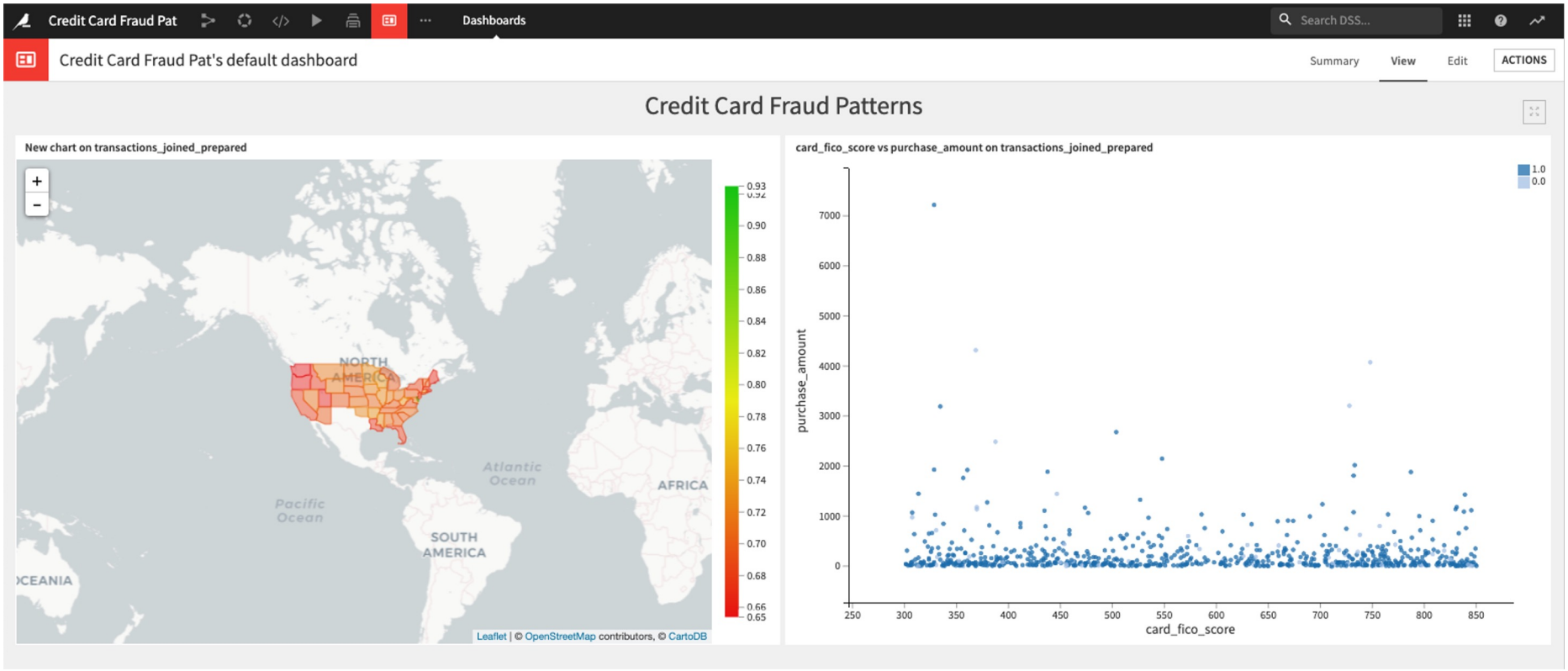



Add more content to your dashboard by clicking the green plus button

- Add a title (e.g. Credit Card Fraud Patterns)

# Dashboards


Credit Card Fraud Pat
Dashboards
Search DSS...
Credit Card Fraud Pat's default dashboard
Summary
View
Edit
ACTIONS
Credit Card Fraud Patterns
New chart on transactions_joined_prepared
NORTH AMERICA
Atlantic Ocean
Pacific Ocean
AFRICA
SOUTH AMERICA
OCEANIA
0.93
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.65
Leaflet | © OpenStreetMap contributors, © CartoDB
card_fico_score vs purchase_amount on transactions_joined_prepared
1.0
0.0
purchase_amount
7000
6000
5000
4000
3000
2000
1000
0
250
300
350
400
450
500
550
600
650
700
750
800
850
card_fico_score