

Exploratory Data Analysis with Dataiku

by Rizki Fajar Nugroho

(date of delivery)

Trainer Profile

Rizki Fajar Nugroho
Data Scientist at SaaS Company
LinkedIn - Rizki Fajar Nugroho



Table of Content

Content
Dataiku 101
Dataiku Set-Up
Data Connection in Dataiku
Exploratory Data Analysis Fundamental and Hands-on

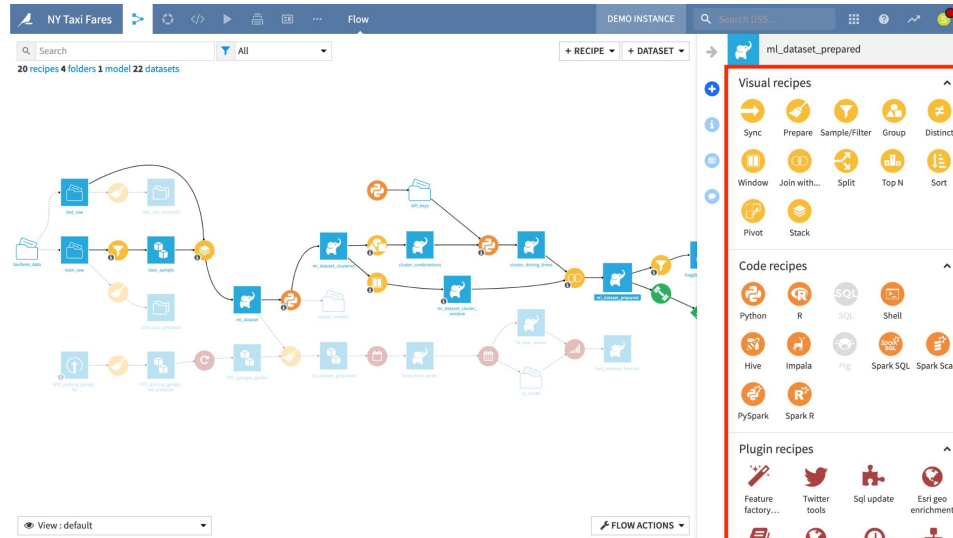




Dataiku 101

What is Dataiku?

Dataiku is a **collaborative** data science platform that empowers company to unlock the full potential of their data. It provides a unified and collaborative environment for data engineers, data scientists, data analysts on data projects.



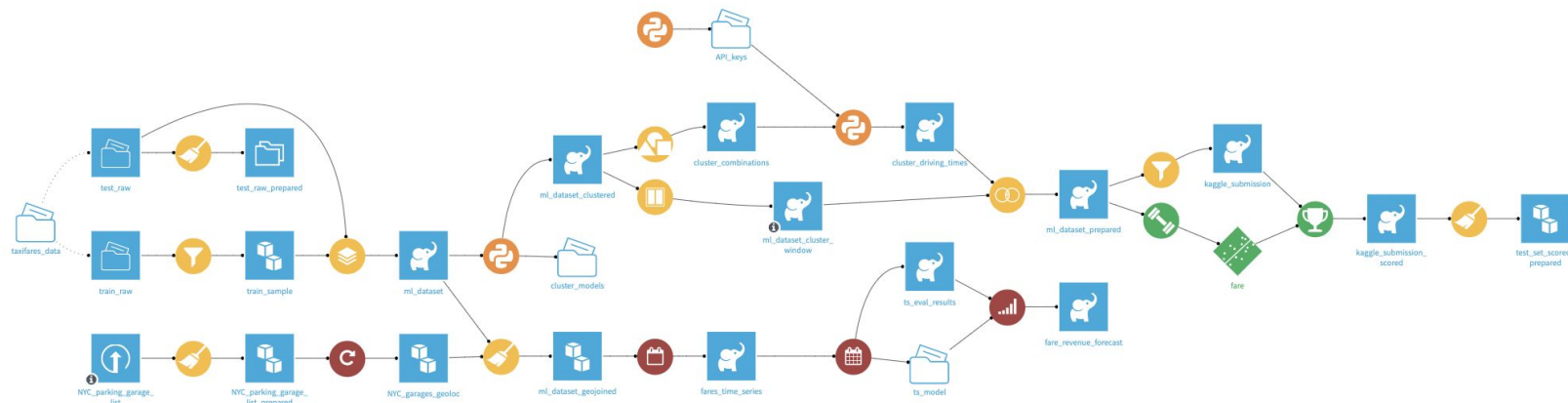
Why Dataiku?

1. **Data Integration:** Seamless integration with diverse data sources (such as SQL database, Amazon S3, and Google Cloud Storage).
2. **Data Preparation:** Intuitive data wrangling and transformation capabilities.
3. **Machine Learning:** Robust support for building, evaluating, and deploying machine learning models.

Why Dataiku?

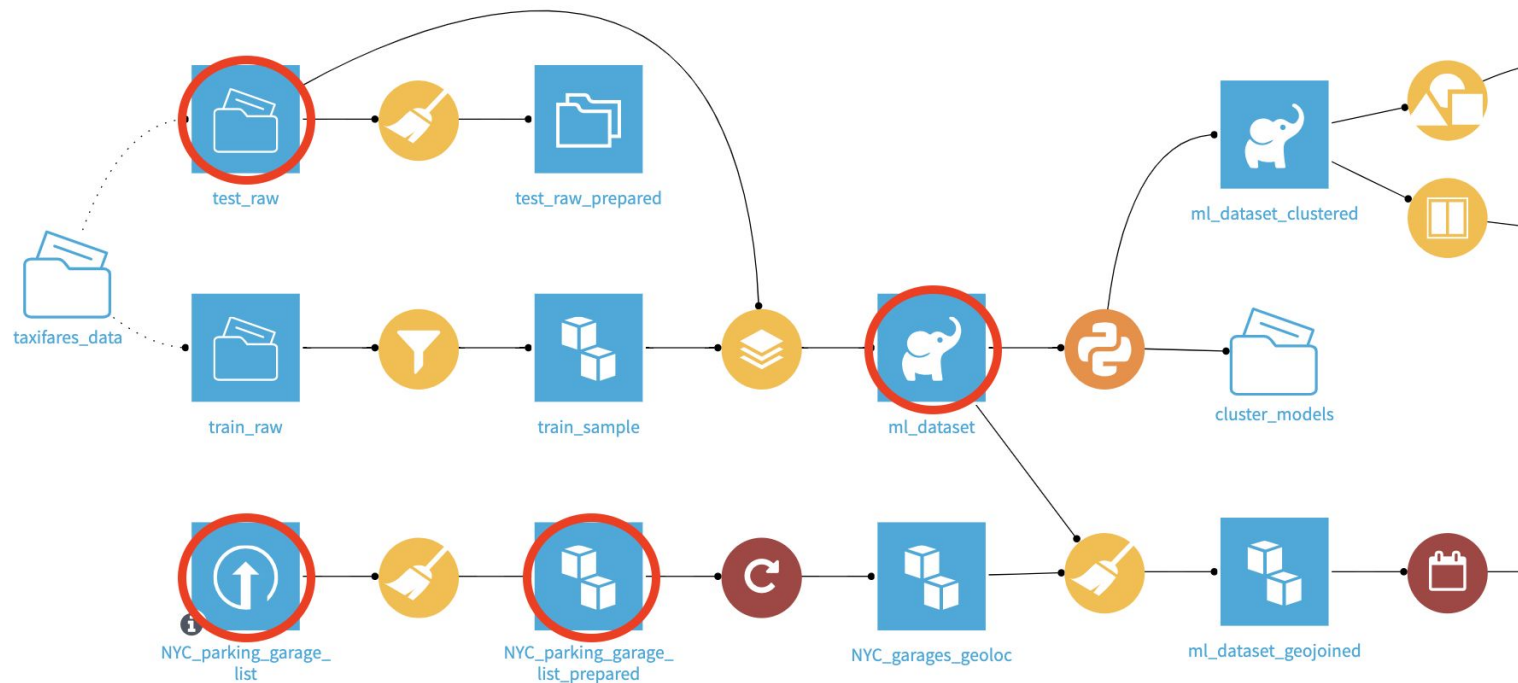
4. **Visual Data Exploration:** User-friendly tools for data visualization and exploration.
5. **Collaboration and Deployment:** Built-in collaboration features and model deployment options.
6. **Governance and Security:** Dataiku provides built-in governance and security features to ensure data privacy, compliance, and risk management

Dataiku Data Analytics Flow and Components

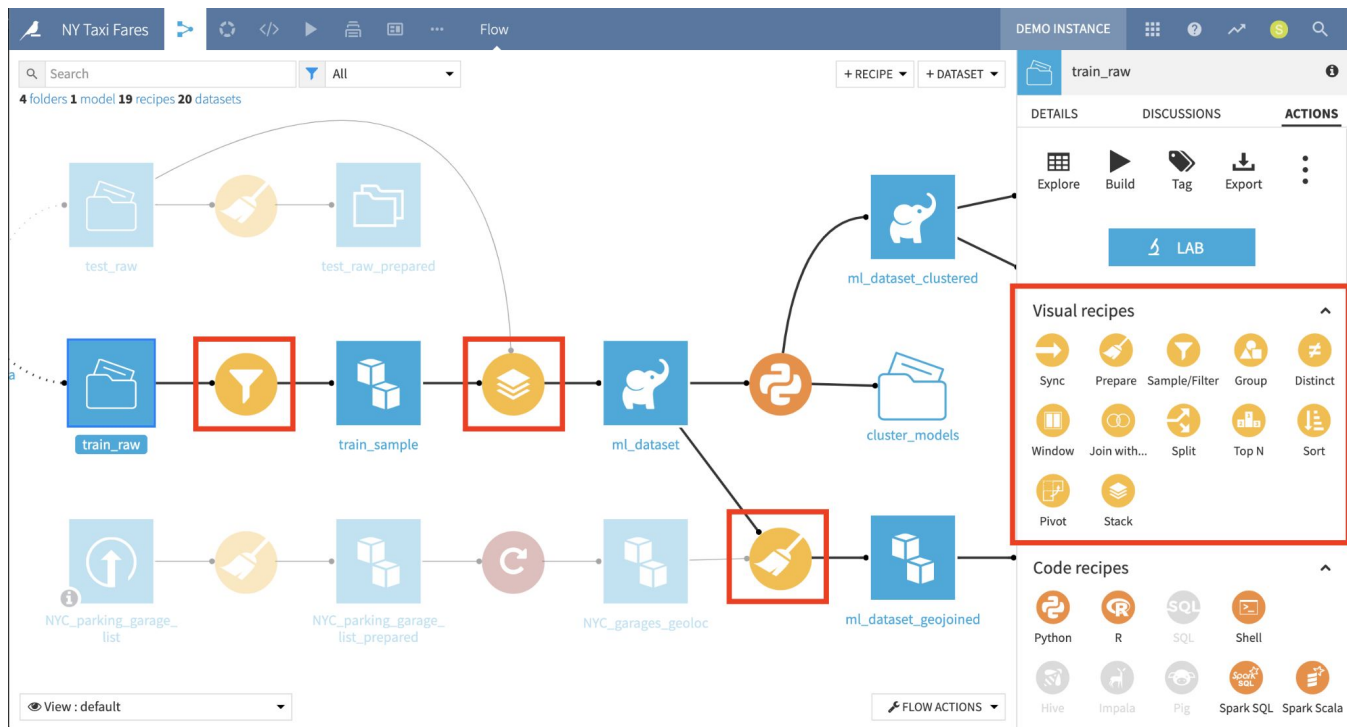


1. Datasets
2. Visual Recipes
3. Code Recipes

Datasets



Visual Recipes



Code Recipes

The screenshot displays the Databricks workspace interface. At the top, a dark blue header bar contains the text 'NY Taxi Fares' and a 'Flow' tab. Below this, a search bar and a filter dropdown set to 'All' are visible. A status bar indicates '4 folders 1 model 19 recipes 20 datasets'. The main area shows a workflow diagram with nodes: 'ml_dataset' (blue elephant icon), 'cluster_models' (blue folder icon), 'ml_dataset_clustered' (blue elephant icon), 'cluster_combinations' (yellow circle with two arrows), 'API_keys' (blue folder icon), 'cluster_driving_times' (blue elephant icon), 'ml_dataset_cluster_window' (blue elephant icon), 'ts_eval_results' (blue elephant icon), and 'cluster_driving_times' (yellow circle with two arrows). Several nodes are highlighted with red boxes: 'API_keys', 'cluster_driving_times', and 'cluster_models'. On the right, a sidebar titled 'cluster_driving_times' has tabs for 'DETAILS', 'DISCUSSIONS', and 'ACTIONS'. The 'ACTIONS' tab is active, showing a 'Code recipes' section with a grid of icons for Python, R, SQL, Shell, Hive, Impala, Pig, Spark SQL, Spark Scala, PySpark, and Spark R. Below this is a 'Plugin recipes' section with icons for Feature, Twitter, Esri geo, Geocoder, Forecast, Named, Rules G..., Sentenc..., Text Su..., and Sentim... At the bottom of the sidebar is an 'Other recipes' section with icons for a flow, a trophy, and a star.

Dataiku Set Up

Prerequisites and Installation Guide

1. Python (Add the python installation to the path)
2. Anaconda
3. Java

Make sure the above programs has been installed and download the dataiku installer from this link - <https://www.dataiku.com/product/get-started/windows/>

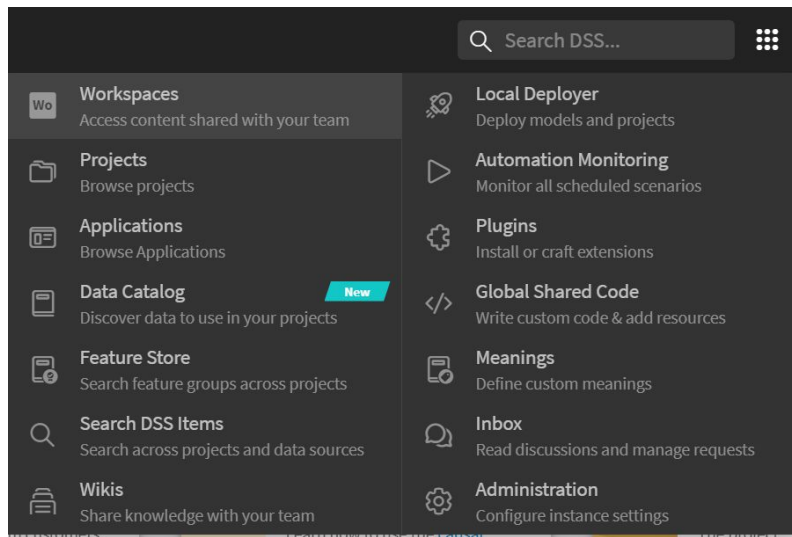
For the installation guide, available from this [link](#)



Data Connection in Dataiku

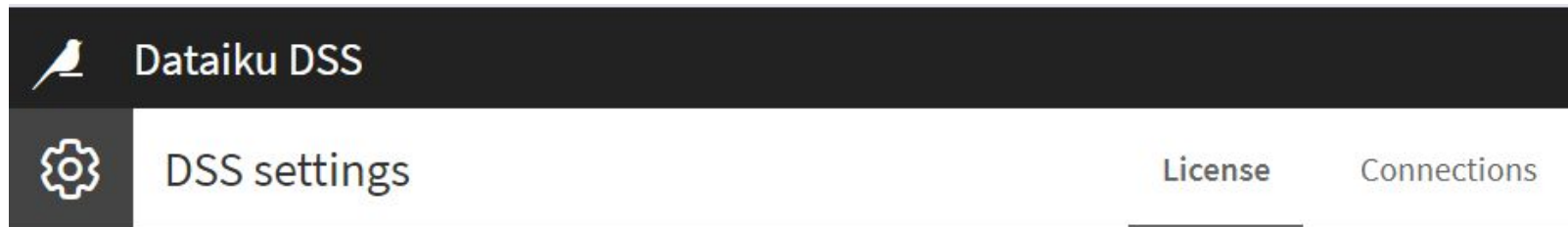
How to Connect your Database to Dataiku?

1. Go to the **Administration**



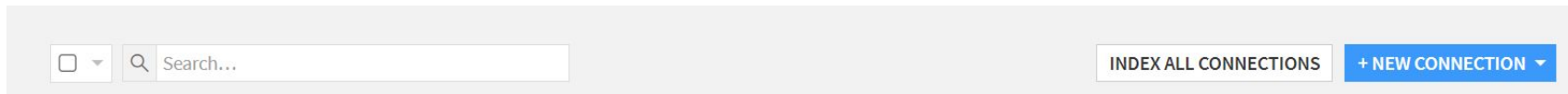
How to Connect your Database to Dataiku?

2. Click on the **Connections**



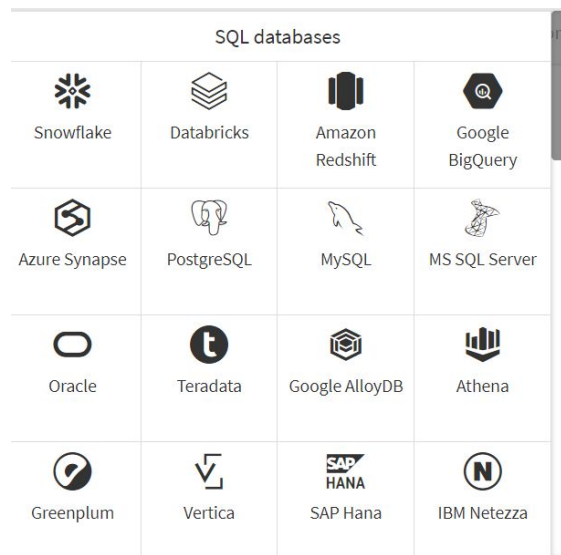
How to Connect your Database to Dataiku?

3. Click on the **new connection** button



How to Connect your Database to Dataiku?

4. Select your respective SQL database





Exploratory Data Analysis Fundamentals

What is Exploratory Data Analysis?

Exploratory Data Analysis refers to the critical process of performing **initial investigations** on data so as to discover **patterns**, to spot **anomalies**, to **check assumption** with the help of **statistical** summary and **graphical** representations

DATA



SORTED



ARRANGED



PRESENTED
VISUALLY



Exploratory Data Analysis Objectives and Scopes

1. Quickly describe a dataset
2. Clean corrupted data
3. Visualize data
4. Calculate and visualize relationships between variables

Data Wrangling

Exploration

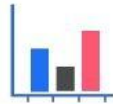
Data Cleansing

Data Wrangling and Cleansing Scopes

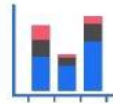
1. Data Retrieval
2. Data Integration
3. Data Quality Report
 - a. Size
 - b. Missing Values Inspection
 - c. Data Type Issues
 - d. Outliers and Anomalies
4. Data Cleansing
 - a. Missing Values Handling
 - b. Data Type Transformation
 - c. Outliers and Anomalies Handling

What is Data Visualization?

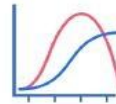
Data visualization is the discipline of trying to **understand data** by placing it in a visual context so that **patterns, trends and correlations** that might not otherwise be detected **can be exposed**.



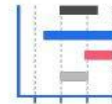
Bar chart



Stacked bar chart



Line graph



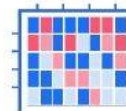
Gantt chart



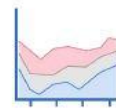
Polar area diagram



Scatter plot



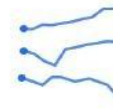
Calendar heatmap



Stacked area chart



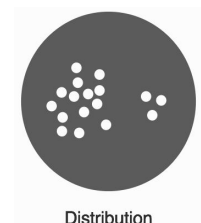
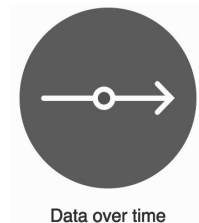
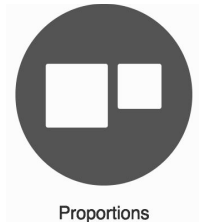
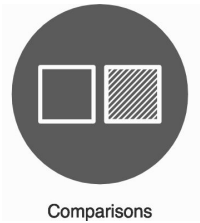
Sparkline



Column sparkline

How to choose the right Data Visualization?

1. Define the purpose of the visualization
2. Define the data you want to display
3. Choose the right representation
4. [References](#)

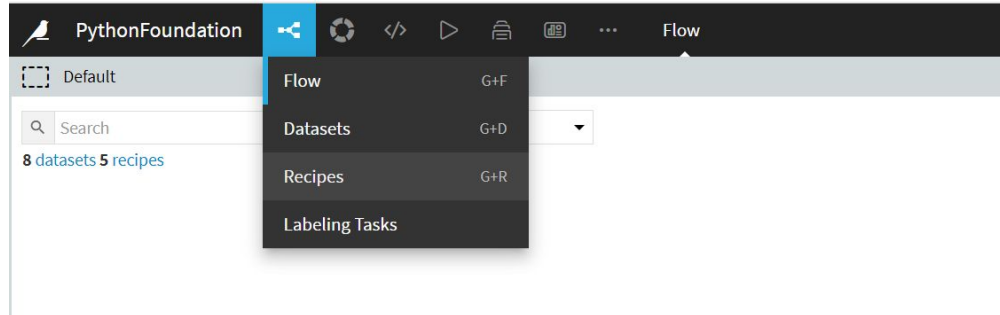




EDA Hands On in Dataiku

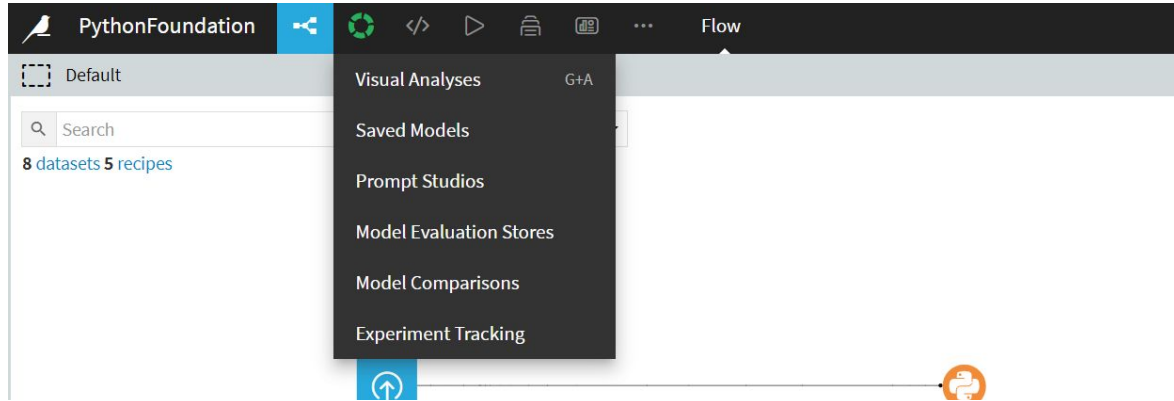
Exploratory Data Analysis Demo - Create Visual Analysis

1. Go to the project workspace in dataiku, choose the flow menu



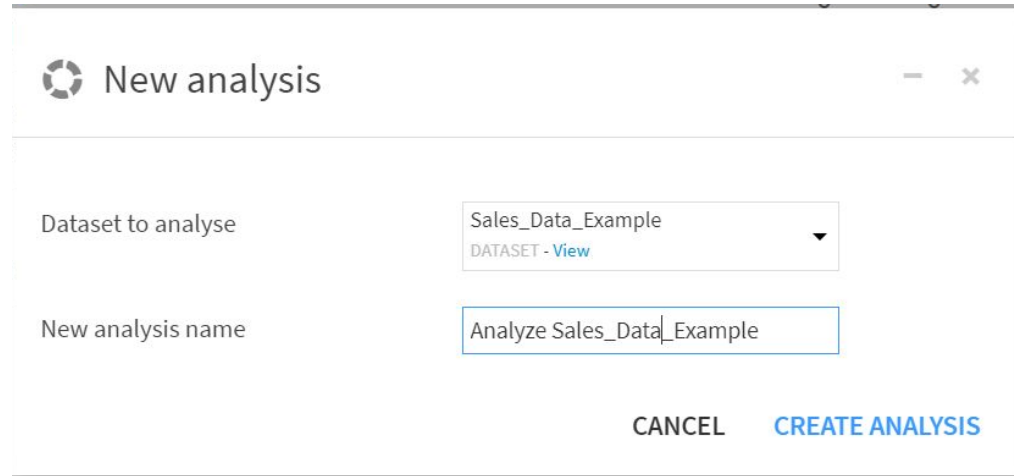
Exploratory Data Analysis Demo - Create Visual Analysis

2. Select the visual analyses section



Exploratory Data Analysis Demo - Create Visual Analysis

3. Add a new analysis and create it with a specific name



The screenshot shows a 'New analysis' dialog box with a title bar containing a refresh icon and the text 'New analysis'. The dialog has two input fields: 'Dataset to analyse' with a dropdown menu showing 'Sales_Data_Example' and a 'DATASET - View' link, and 'New analysis name' with a text input field containing 'Analyze Sales_Data_Example'. At the bottom right, there are two buttons: 'CANCEL' and 'CREATE ANALYSIS'.

New analysis

Dataset to analyse: Sales_Data_Example
DATASET - [View](#)

New analysis name: Analyze Sales_Data_Example

CANCEL **CREATE ANALYSIS**

Exploratory Data Analysis Demo - Create Visual Analysis

4. Select the **script** section and click the *add a new step* button to create and experiment with the required data preprocessing

The screenshot displays the PythonFoundation Visual Analyses interface. The top navigation bar includes the PythonFoundation logo, a share icon, a refresh icon, a code editor icon, a play icon, a document icon, a menu icon, and the text "Visual Analyses". Below the navigation bar, the main title is "Analyze Sales_Data_Example_Rename_Columns". The interface is divided into two main sections: "Script" and "Sample settings". The "Script" section shows a list of steps, with the first step being "Rename column 'OrderYear' to 'YearOfOrder'" (ID: 997). Below the list is a green button labeled "+ ADD A NEW STEP" and a grey button labeled "+ ADD A GROUP". The "Sample settings" section shows "1 step" and "First 10,000 rows". To the right of the "Script" section is a "Step preview on whole data" section, which includes a "Disable preview" button and a preview table. The preview table has columns "OrderID", "CustomerID", "ProductID", and "C" (partially visible). The data rows are:

OrderID	CustomerID	ProductID	C
Integer	Integer	Integer	Int
2	48	81	
3	118	83	
4	102	28	

Exploratory Data Analysis Demo - Create Visual Analysis

- Once the required steps are made, deploy the script to change it into a recipe in Dataiku by clicking the *deploy script* button

The screenshot displays the Dataiku interface for a script step named "Analyze Sales_Data_Example_Rename_Columns". The interface includes a "Script" tab, "Sample settings" (First 10,000 rows), and a "Deploy Script" button. A step preview shows the action "Rename column 'OrderYear' to 'YearOfOrder'" with 997 rows affected. Below the preview, a table displays sample data for 997 rows.

OrderID	CustomerID	ProductID	Quantity	Price	OrderDate	Age	Gender
Integer	Integer	Integer	Integer	Decimal	Date (unparsed)	Integer	Gender
2	48	81	2	414.1818294989762	2020-05-04 18:42:53	22	Other

Dataiku - Data Size Report

The screenshot shows the Dataiku interface for a dataset named 'Sales_Data_Example'. The top navigation bar includes 'Explore', 'Charts', 'Statistics', 'Status', 'History', 'Settings', and 'ACTIONS'. Below the navigation bar, there is a 'Whole data' button indicating '1,000 rows'. To the right, there are buttons for 'DISPLAY', 'TABLE', 'COLUMNS', and a chart icon. At the bottom right, a status bar indicates '1,000 rows - 16 columns'.

There are also key attributes of a **Datasets** such as:

1. shows dimensionality of the DataFrame
2. shows the **Datasets** column name

The screenshot shows the Dataiku interface for a dataset named 'Sales_Data_Example'. The top navigation bar includes 'Explore', 'Charts', 'Statistics', 'Status', 'History', 'Settings', and 'ACTIONS'. Below the navigation bar, there is a 'Whole data' button indicating '1,000 rows'. To the right, there are buttons for 'Metrics on', 'Sample', 'Index', 'TABLE', and 'COLUMNS'. Below these buttons, there is a table listing the dataset's attributes.

<input type="checkbox"/> 0 / 16	Filter	All meanings	% valid
<input type="checkbox"/> OrderID	Integer	100.00%	
<input type="checkbox"/> CustomerID	Integer	100.00%	
<input type="checkbox"/> ProductID	Integer	100.00%	
<input type="checkbox"/> Quantity	Integer	100.00%	
<input type="checkbox"/> Price	Decimal	100.00%	
<input type="checkbox"/> OrderDate	Date (Unparsed)	100.00%	
<input type="checkbox"/> Name	Text	100.00%	

Dataiku - Data Exploration - Statistics Report

▼ # Quantity	▼ # Price	▼ # Age	▼ # UnitPrice
> Histogram	> Histogram	> Histogram	> Histogram
▼ Summary stats	▼ Summary stats	▼ Summary stats	▼ Summary stats
N values1000	N values1000	N values1000	N values1000
N distinct9	N distinct1000	N distinct58	N distinct100
N finite1000	N finite1000	N finite1000	N finite1000
Mean5.002	Mean250.97785534	Mean42.463	Mean243.97882996
Median5	Median253.39693612	Median43	Median249.80776817
Std Dev2.5648732083	Std Dev150.49041309	Std Dev16.866508533	Std Dev141.30288149
Min1	Min-445.7081802	Min-57	Min15.951221725
Max9	Max499.71895226	Max69	Max484.23346737

Dataiku - Data Exploration - Statistics Report

▼ Gender



> Histogram

▼ Summary stats

N values	1000
N distinct	3
Mode	Female
N empty	0

▼ Frequency table

Female	38%	377
Other	33%	328
Male	30%	295
N distinct		3

▼ Category



> Histogram

▼ Summary stats

N values	1000
N distinct	4
Mode	Clothing
N empty	0

▼ Frequency table

Clothing	29%	285
Sports	28%	283
Electronics	26%	261
Home & Kitchen	17%	171
N distinct		4

▼ Location



> Histogram

▼ Summary stats

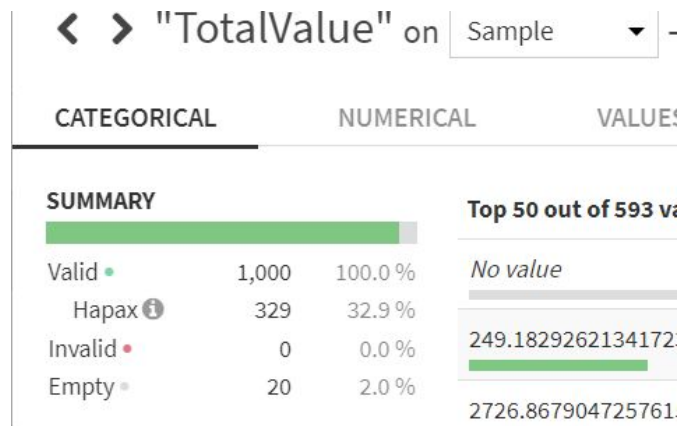
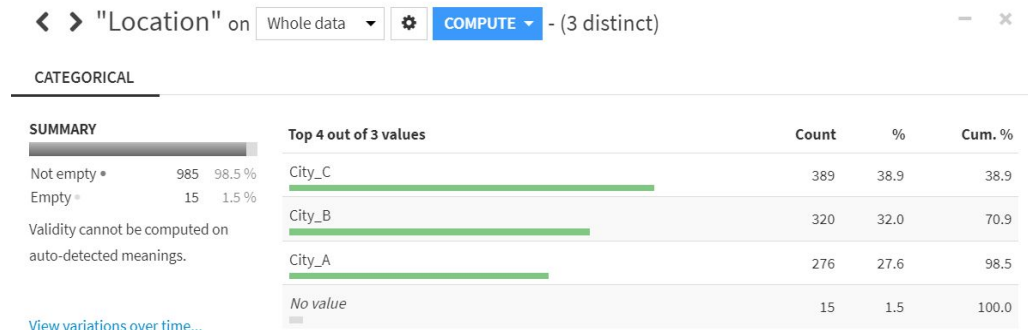
N values	1000
N distinct	4
Mode	City_C
N empty	15

▼ Frequency table

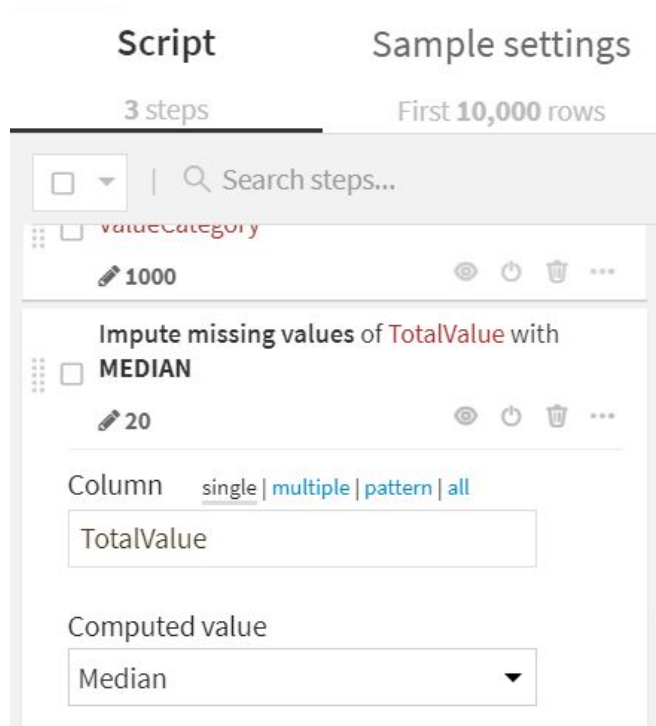
City_C	39%	389
City_B	32%	320
City_A	28%	276
(no value)	2%	15
N distinct		4

Dataiku - Missing Value Handling

Click the analyze section on the specific column name for missing values inspection







Dataiku - Missing Value Handling





- For example, impute the missing value record on the **TotalValue** column with the median value
- There are other options to impute the missing value either by mean, median, or mode

Dataiku - Anomalies and Outlier Handling

☐ Clip values > 1500 in Price



Columns

 Price 

[+ ADD A COLUMN](#)

Lower bound

Upper bound

☒ Clip outliers

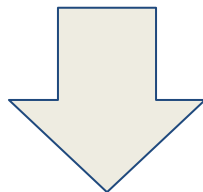
☐ Clear outliers

In the script section, there are two choices available,

1. Clip outliers / anomalies record by a certain value
2. Clear / remove the outliers / anomalies record if it's exceeding the threshold

Dataiku - Data Type Inspection and Transformation

City	Price	OrderDate	Name	Age	Gender	Location	ProductName	Category
	string Decimal	string Date (unparsed)	string Text	string Integer	string Gender	string Text	string Text	string Text
2	414.1818294989762	2020-05-04 18:42:53	Customer_48	22	Other	City_B	Product_81	Electronics
8	147.74873927279793	2020-06-08 10:57:29	Customer_118	43	Male	City_B	Product_83	Clothing
9	239.4031566884592	2020-02-03 00:42:43	Customer_193	30	Female	City_C	Product_38	Clothing
4	414.6722037546345	2020-08-09 16:56:32	Customer_252	65	Other	City_C	Product_95	Electronics



Price	OrderDate	Name	Age	Gender	Location	ProductName	Category
float Decimal	date Date (unparsed)	string Text	int Integer	string Gender	string Text	string Text	string Text

Exploratory Data Analysis Demo in Dataiku

Let's head to the dataiku and try it out

Thank you!