IYKRA
Data Fellowship Program

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

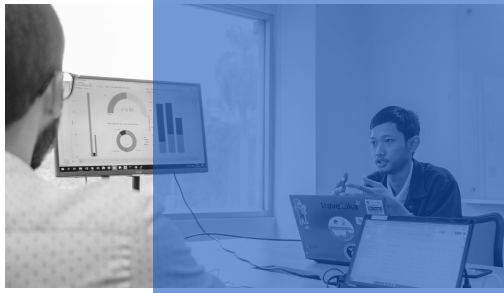
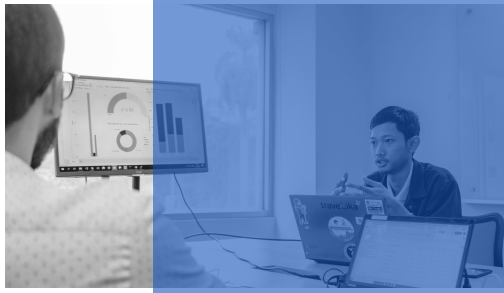


Table of Content

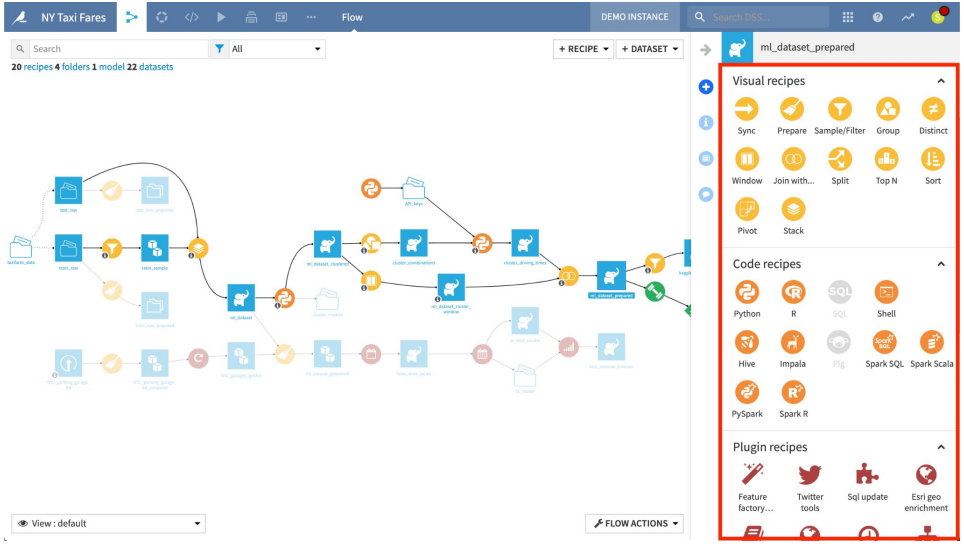


Data Fellowship

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.



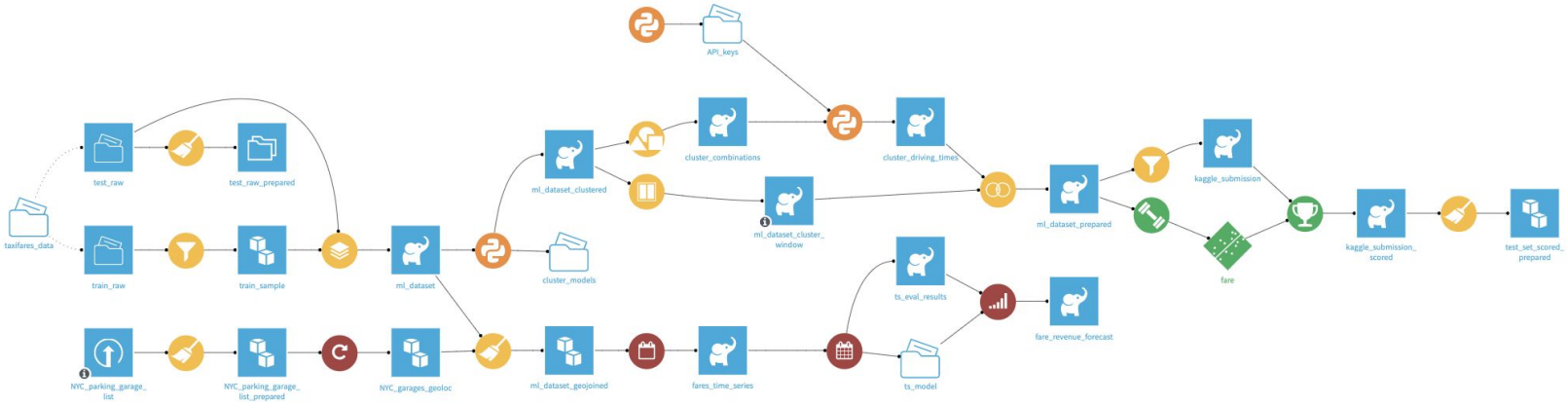
Why Dataiku?

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?

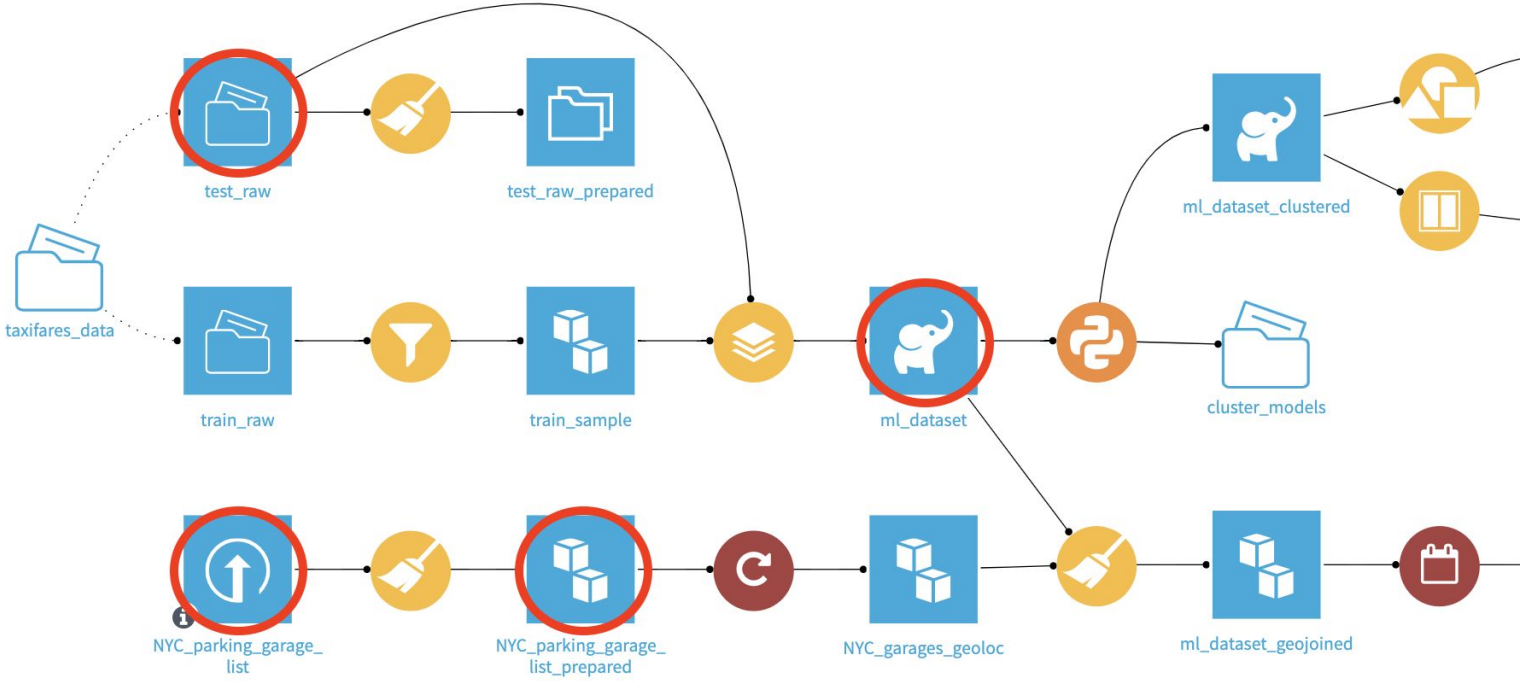
4. **Visual Data Exploration**: User-friendly tools for data visualization and exploration.
5. **Collaboration and Deployment**: Built-in collaboration features and model deployment options.
6. **Governance and Security:** Dataiku provides built-in governance and security features to ensure data privacy,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

Dataiku Data Analytics Flow and Components

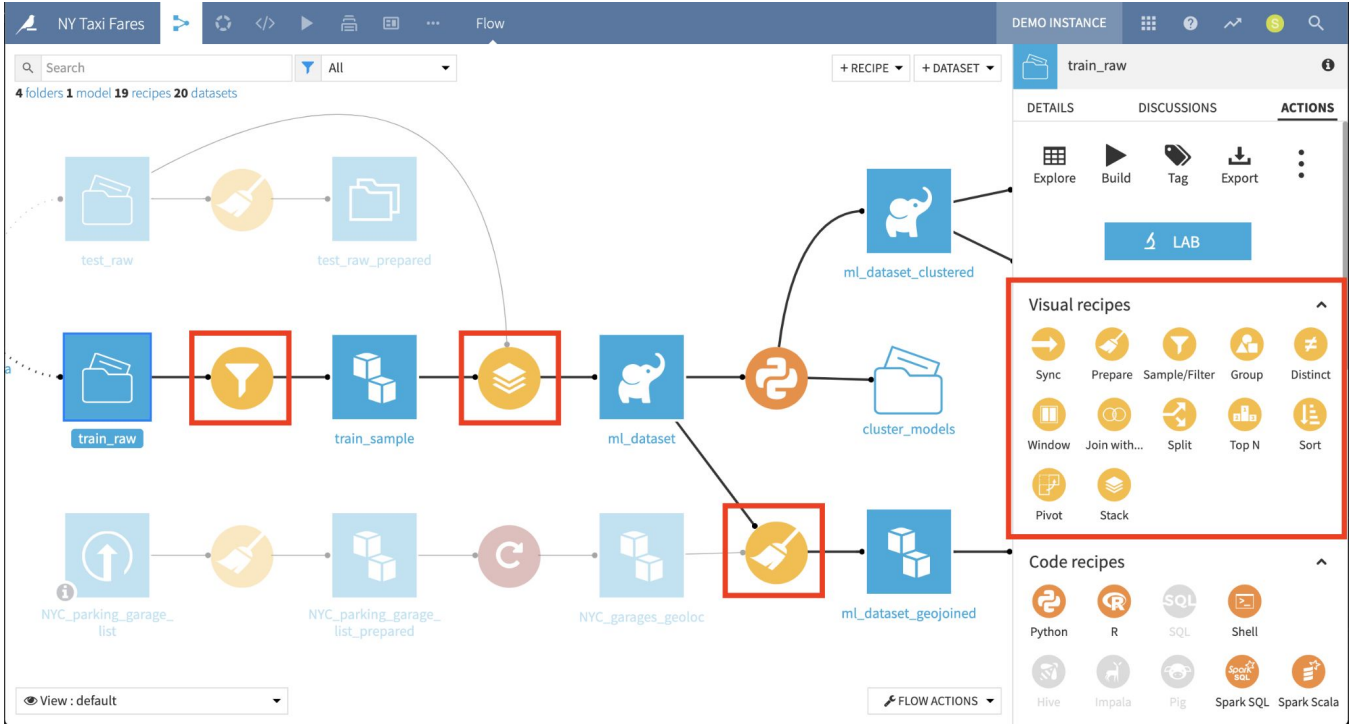


1. Datasets
2. Visual Recipes
3. Code Recipes

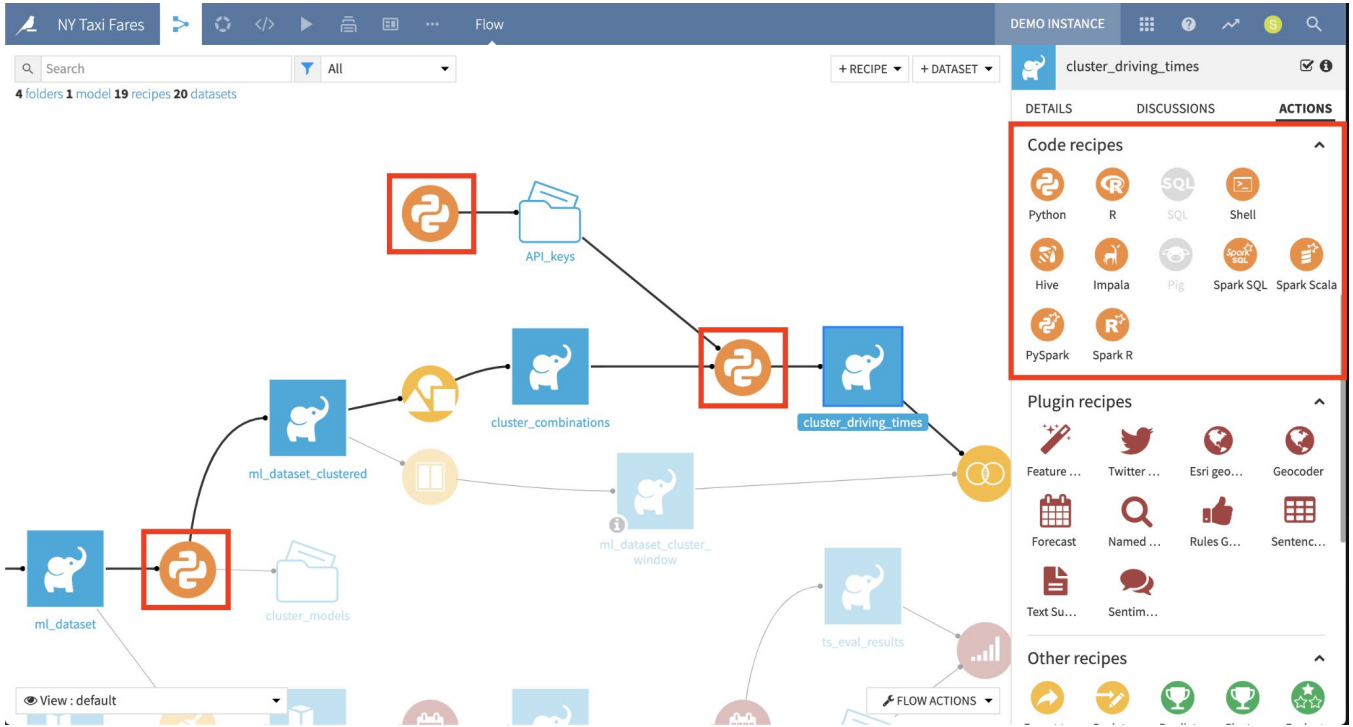
Datasets


test_raw
test_raw_prepared
taxifares_data
train_raw
train_sample
ml_dataset
ml_dataset_clustered
cluster_models
NYC_parking_garage_
list
NYC_parking_garage_
list_prepared
NYC_garages_geoloc
ml_dataset_geojoined

Visual Recipes



Code Recipes



Data Fellowship

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/

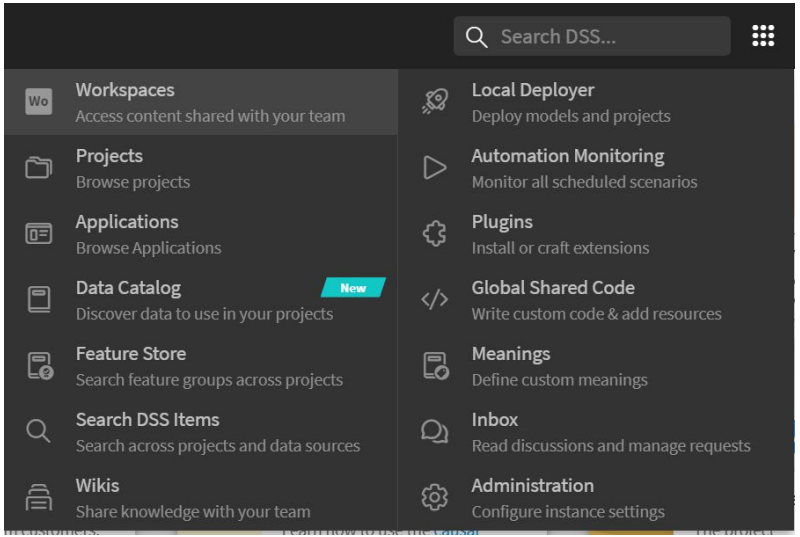
For the installation guide, available from this link

Data Fellowship

Data Connection in Dataiku

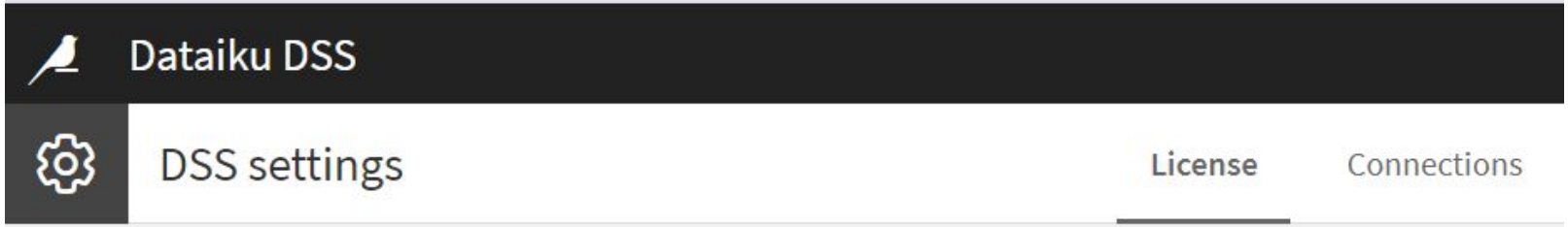
How to Connect your Database to Dataiku?

1. Go to the **Administration**



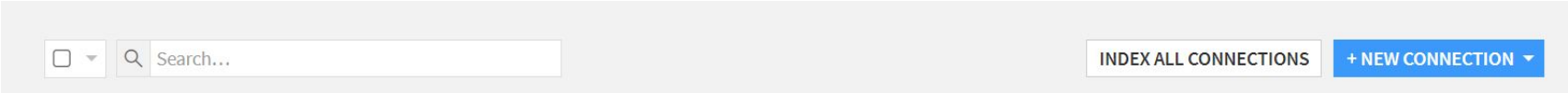
How to Connect your Database to Dataiku?

2. Click on the **Connections**



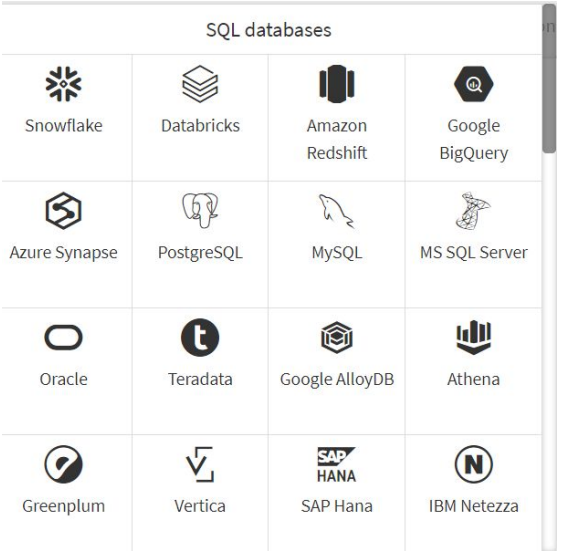
How to Connect your Database to Dataiku?

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



How to Connect your Database to Dataiku?

4. Select your respective SQL database

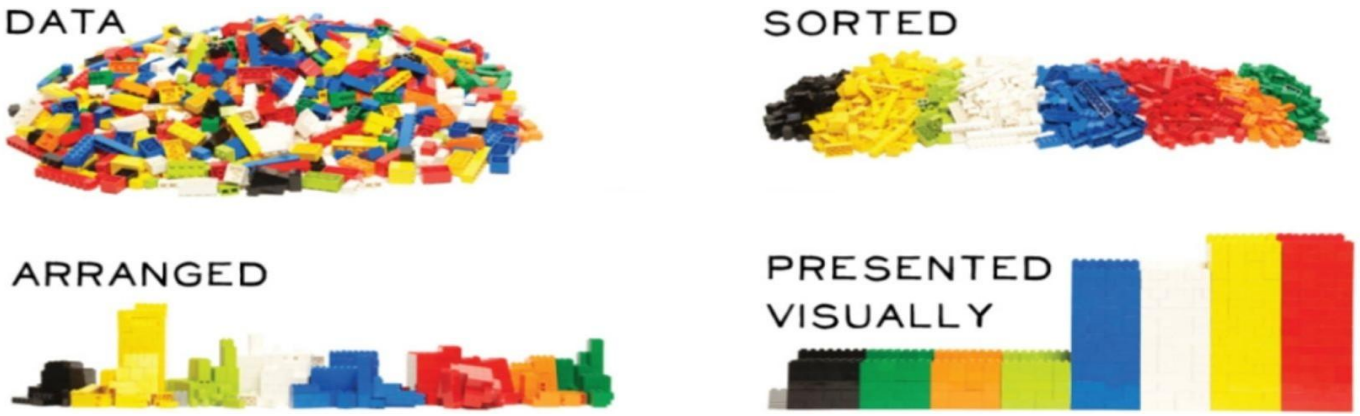


Data Fellowship

Exploratory Data Analysis Fundamentals

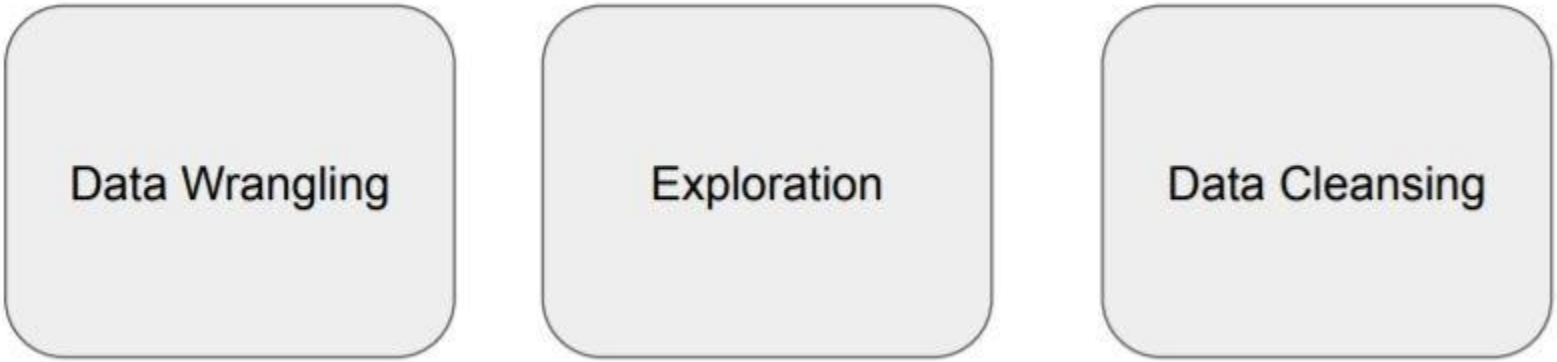
What is Exploratory Data Analysis?

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 of **statistical** summary and **graphical** representations



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

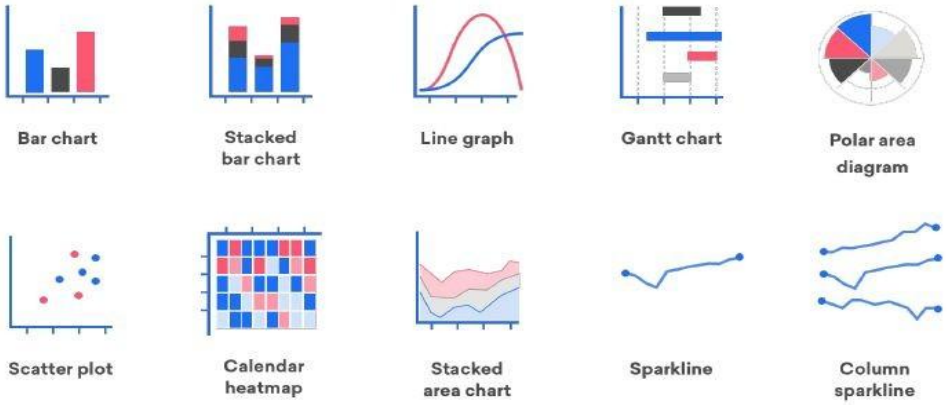


Data Wrangling and Cleansing Scopes

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

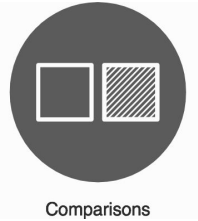
What is Data Visualization?

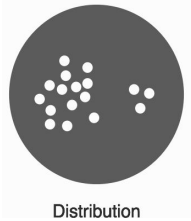
Data visualization is the discipline of trying to **understand data** by placing it in a visual context so that **patterns, trends and correlations** that might not otherwise be detected **can be exposed**.

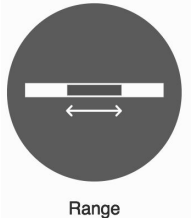


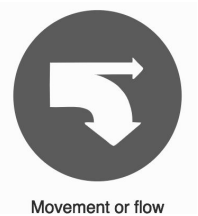
How to choose the right Data Visualization?

1. Define the purpose of the visualization
2. Define the data you want to display
3. Choose the right representation
4. References







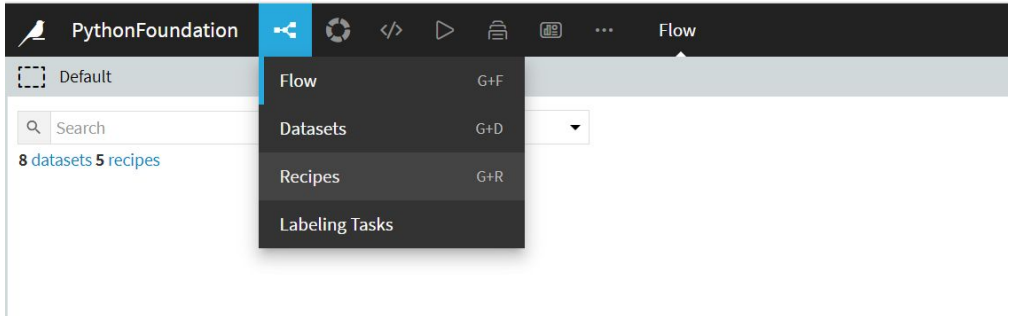


Data Fellowship

EDA Hands On in Dataiku

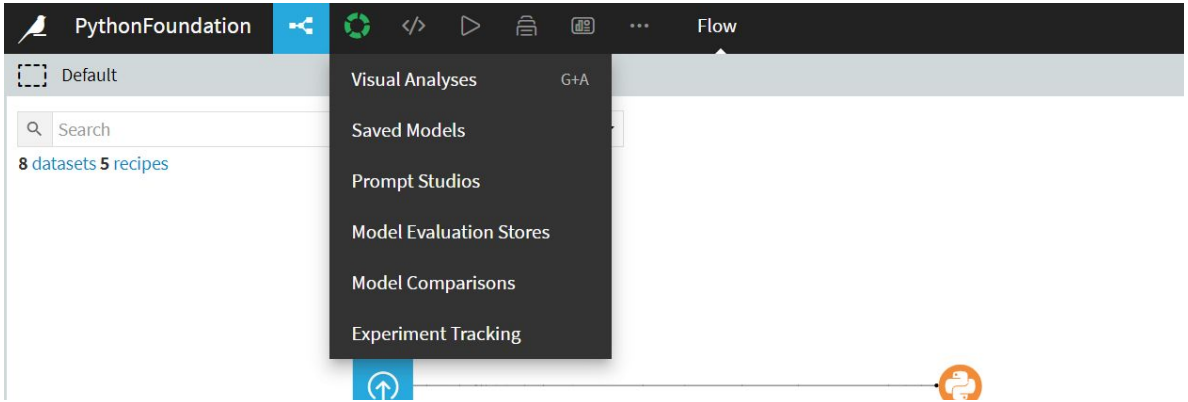
Exploratory Data Analysis Demo - Create Visual Analysis

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



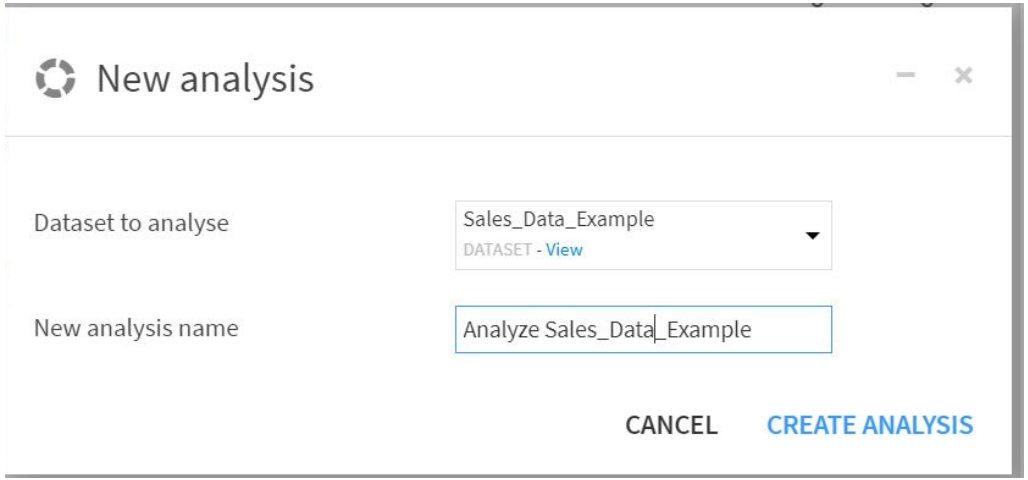
Exploratory Data Analysis Demo - Create Visual Analysis

2. Select the visual analyses section



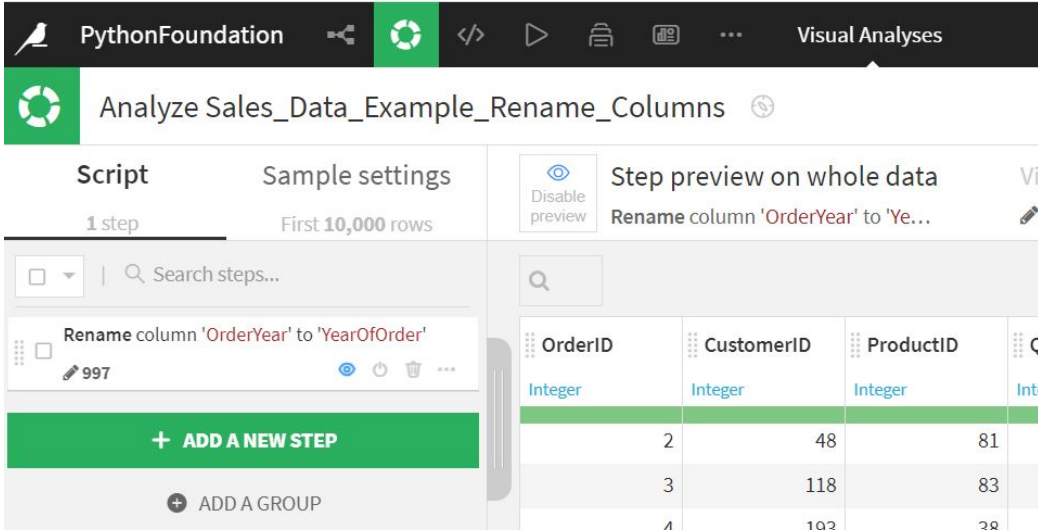
Exploratory Data Analysis Demo - Create Visual Analysis

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



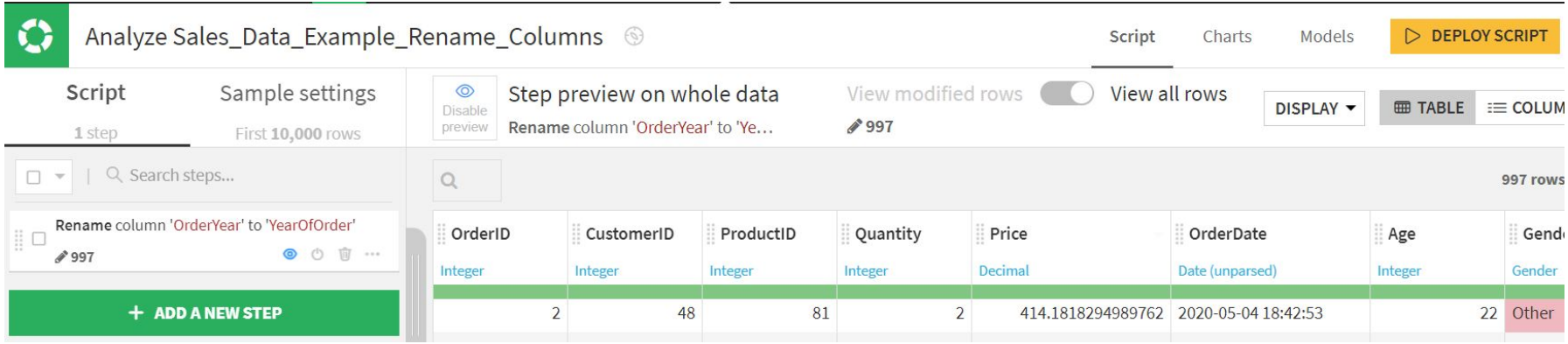
Exploratory Data Analysis Demo - Create Visual Analysis

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

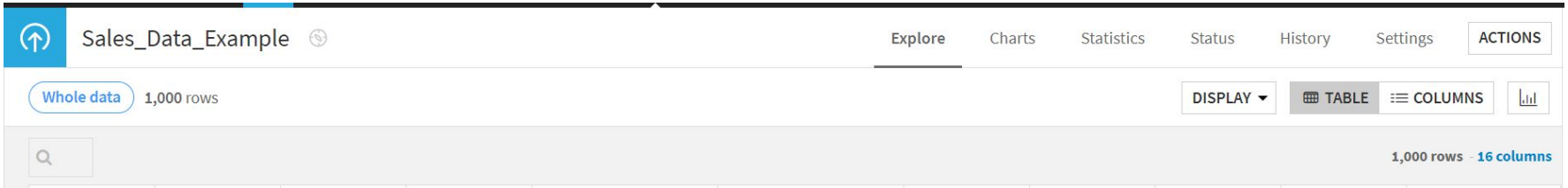


Exploratory Data Analysis Demo - Create Visual Analysis

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

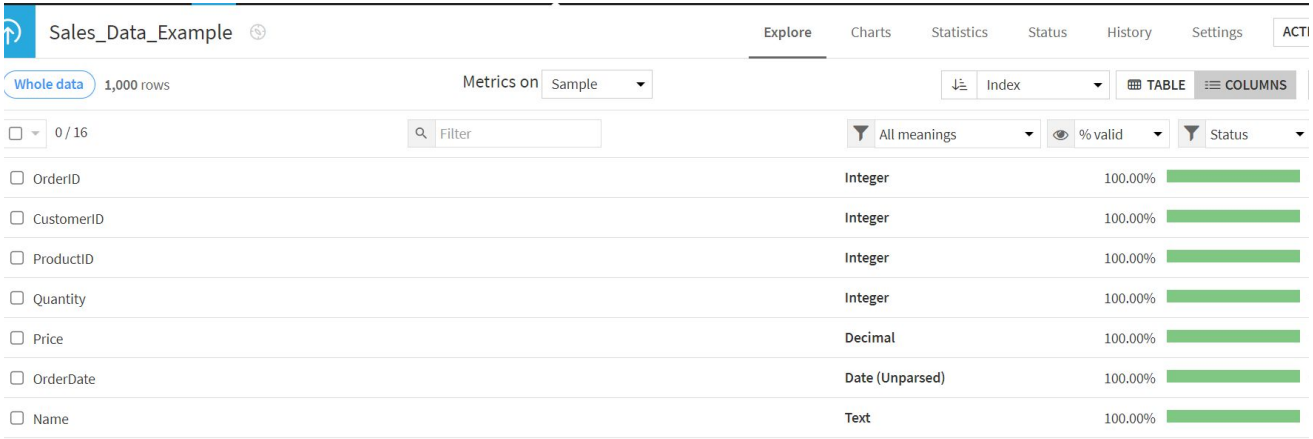


Dataiku - Data Size Report



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

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



Dataiku - Data Exploration - Statistics Report

Quantity

Histogram

Summary stats

N values	1000
N distinct	9
N finite	1000
Mean	5.002
Median	5
Std Dev	2.5648732083
Min	1
Max	9

Price

Histogram

Summary stats

N values	1000
N distinct	1000
N finite	1000
Mean	250.97785534
Median	253.39693612
Std Dev	150.49041309
Min	-445.7081802
Max	499.71895226

Age

Histogram

Summary stats

N values	1000
N distinct	58
N finite	1000
Mean	42.463
Median	43
Std Dev	16.866508533
Min	-57
Max	69

UnitPrice

Histogram

Summary stats

N values	1000
N distinct	100
N finite	1000
Mean	243.97882996
Median	249.80776817
Std Dev	141.30288149
Min	15.951221725
Max	484.23346737

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

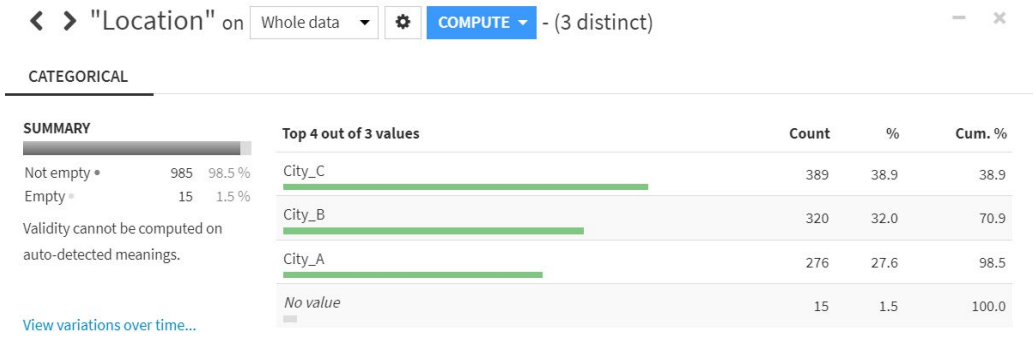
N values	1000
N distinct	4
Mode	City_C
N empty	15

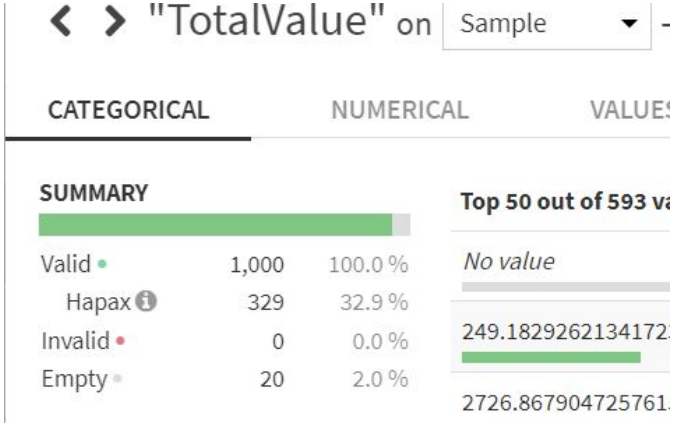
Frequency table

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

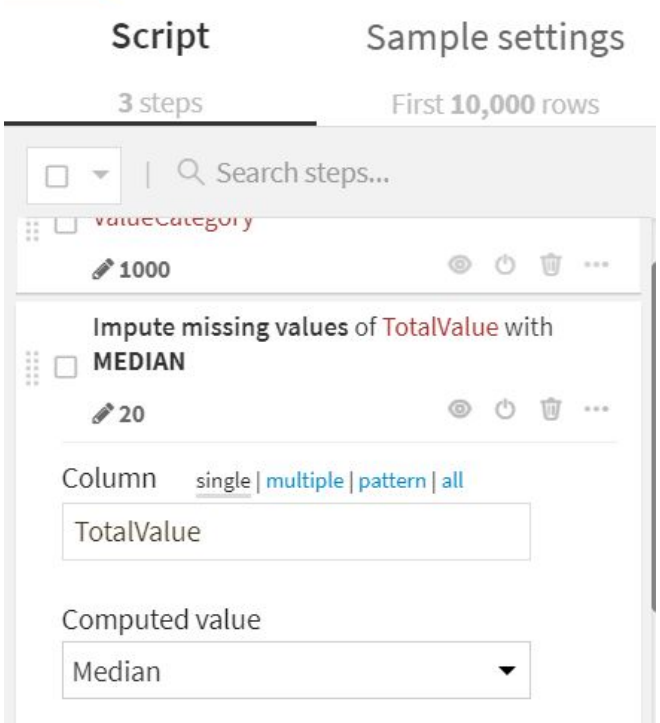
Dataiku - Missing Value Handling

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



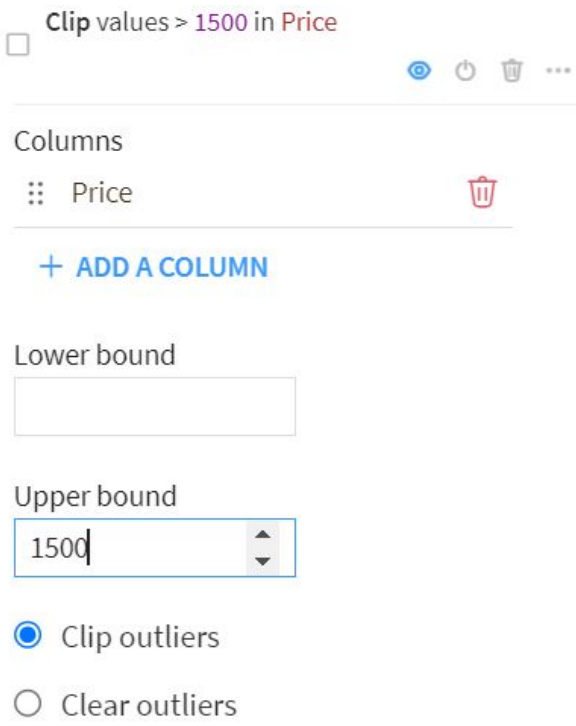


Dataiku - Missing Value Handling



- 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



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

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

Exploratory Data Analysis Demo in Dataiku

Let’s head to the dataiku and try it out

iykra Data Fellowship

Thank you!