**Step by Step on r/place data analysis via Dataiku V1.**

**Team:**

Mehrpad Moattari (Narrator, please reach out to him for comments or concerns on this file).

Jacob Shoenburg

Olliver Sheehan

Parker Peterson

**Data Upload:**

CSV files were uploaded by 2 members of the team to their local operating systems and uploaded to Dataiku using the upload feature.

For more information on upload steps please visit the following link and choose the Core Designer course under Learning Paths.

<https://academy.dataiku.com/>

**Overview of the entire ‘flow’ on Dataiku and legend for the steps.**

A diagram of a network

Description automatically generated

**Step 5**

**Step 6.C**

**Step 6.B**

**Step 6.A**

**Step 6**

**Step 5.A**

**Step 4.D**

**Step 4.C**

**Step 4.B**

**Step 4.A**

**Step 4**

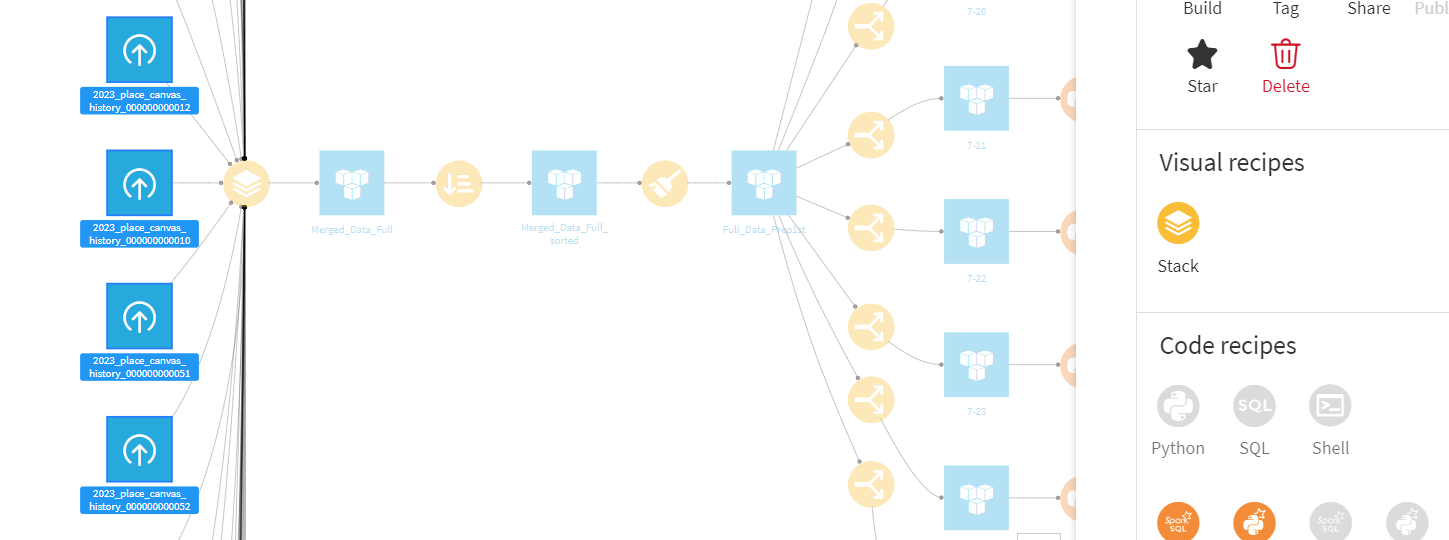
**Step 3**

**Step 2**

**Step 1**

Step1:

Select all the datasets and use the Stack action to stack all of the datasets into one large datasets.



Step2:

Once all the datasets are merged, use the Sort action to sort the data from the smallest to the largest timestamp.

A diagram of a blockchain

Description automatically generated

A screenshot of a computer

Description automatically generated

Step3:

After the Sort action, select the Prepare action shown below to turn the Timestamp into two new columns Date and Time.

A diagram of a blockchain

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Step4:

After the creation of the two new columns, use the Group action to group the data by the column ‘user’. As shown below:

A screenshot of a phone

Description automatically generated

A screenshot of a computer

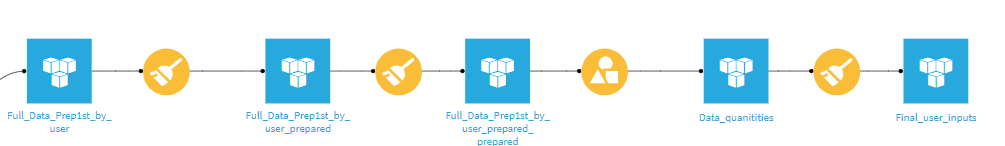
Description automatically generated

A screenshot of a computer

Description automatically generated

Step4A:

Use the Prepare action to rename ‘count’ column to ‘quant’.



Step4b:

Use the Prepare action to create the ‘Quant Range’ column out of the ‘quant’ column.

A blue square with white logo

Description automatically generated

A screenshot of a computer

Description automatically generated

Step4c:

A blue square with white logo

Description automatically generated

Use the Group action to group the “Quant Range” based on the user counts.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Step4d:

Use the Prepare action to get a table that outputs the number of users per range pixels submission.

A blue square with white logo

Description automatically generated

A screenshot of a computer

Description automatically generated

Step5:

Using the Prepare action to perform parsing as shown below for further analysis on the hourly inputs for users.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Step5A:

Use the Group by action to group the data by Day and Hour.

A long thin black line

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

Step6:

Use the Split action to split the data into each separate day from 7-20 to 7-25.

A screenshot of a computer

Description automatically generated

Step6A:

Once the data is separated into each day, run the following Python Code using the Python action to indicate the existence of Bot for the entry.

A diagram of a computer program

Description automatically generated with medium confidence

A screenshot of a phone

Description automatically generated

The Python script:

# -\*- coding: utf-8 -\*-

import dataiku

import pandas as pd, numpy as np

from dataiku import pandasutils as pdu

input\_dataset = dataiku.Dataset("7-20")

output\_dataset = dataiku.Dataset("7-20\_bots")

with output\_dataset .get\_writer() as writer:

for input\_chunk\_df in input\_dataset.iter\_dataframes(10000):

# input\_chunk\_df is a dataframe containing just a

# chunk of data from input\_dataset, with at most 10000 records

df = input\_chunk\_df

# Process here the data ...

df['isBot'] = df.groupby(['user', 'coordinate', 'pixel\_color'])['user'].transform('size').ge(2).astype(int)

df.loc[df['isBot'] == 1, 'isBot'] = 1

output\_chunk\_df = df

# Append the processed chunk to the output

writer.write\_dataframe(output\_chunk\_df)

Step6B:

After the Python script is completed for each day, stack the data similar to Step1.

A diagram of a blockchain

Description automatically generated

Step6C:

After the data is merged again, use the Sample/Filter Recipe to filter for bots.

A screenshot of a computer

Description automatically generated

Some of the outputs were exported as small Excel sheets and charts were made via Jypyter notebook, and some charts were made within Dataiku by using the Charts feature that’s available to each dataset in the Flow. Further analysis could be done by using the Statistics feature.

A screenshot of a computer

Description automatically generated