**Assignment 1: Data Cleaning and Observation**

COS10022 – Introduction to Data Science

Semester 1, 2022

Names: Ngo Cong Thanh, Nguyen Bach, Hoang Khanh Bang(Group 9)

Student ID: 103433609, 103581540, 103565816

Date: 21/02/ 2022

1. Observed Information:

|  |  |  |  |
| --- | --- | --- | --- |
| **Question** | **Answer** | **Question** | **Answer** |
|  | 6150 |  | 5861 |
|  | 5778 |  | 164 |
|  | 1078 |  | 5566 |
|  | 5023 |  | 229 |
|  | 2 |  | 1 |
|  | 5698 |  | 4388 |
|  | 266 |  | 800 |
|  | 15 |  | 0 |
|  | 2 |  | 5 |
|  | 13 |  | 4 |
|  | 250 |  | 9 |
|  | 5861 |  | 4.699% |
|  | Callsign |  | Country (some attribute is not a country) |

1. [Put your answers here.]

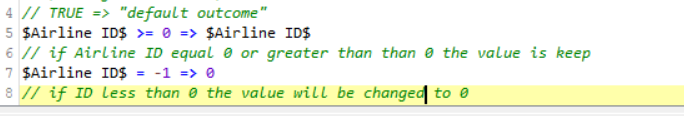
The country which owns the most active operational routes is United State with 149 routes active.

1. [Screenshot of KNIME workflow or codes with comments.]

Ảnh có chứa bàn

Mô tả được tạo tự động



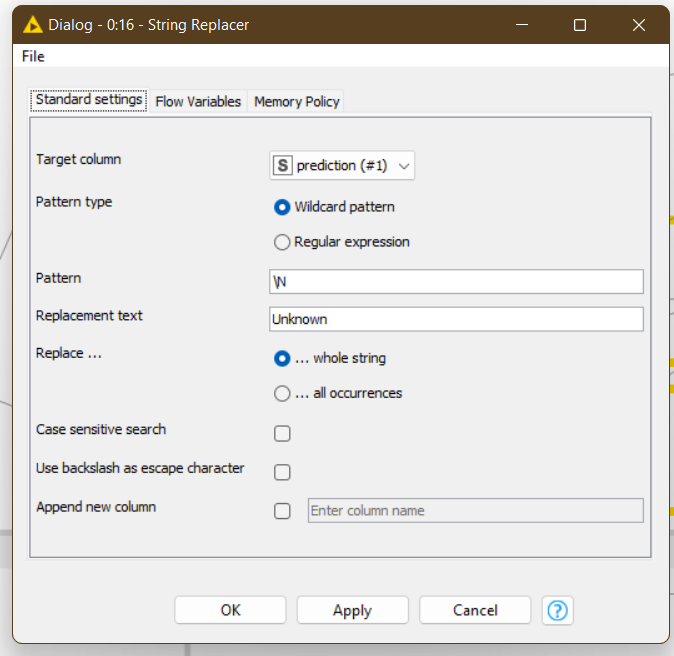




Ảnh có chứa văn bản

Mô tả được tạo tự động



Ảnh có chứa văn bản

Mô tả được tạo tự động

Every \N change to Unknown and every alias contain English letter and number is allowed to keep



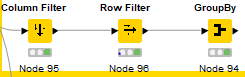








For Question 10 to 16, our team use the same model to find out the answer:



Ảnh có chứa bàn

Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

14)

Ảnh có chứa bàn

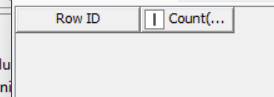
Mô tả được tạo tự động

15)

Ảnh có chứa bàn

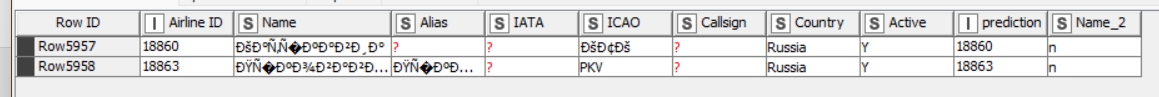
Mô tả được tạo tự động

16)

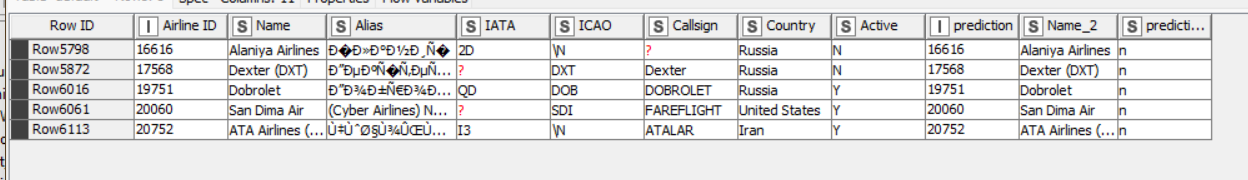


For question 17 to 23, the **Joiner** node  was used to sort out all of the cleared elements of each column.

17)



18)

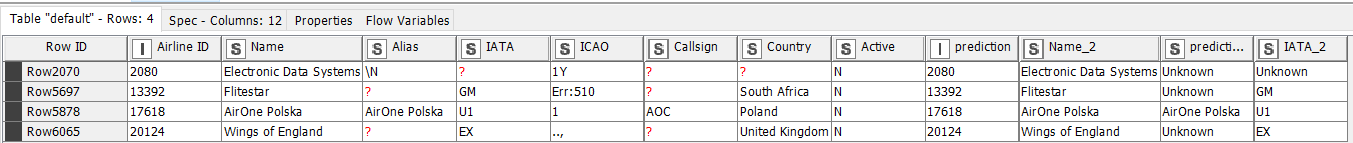


19)

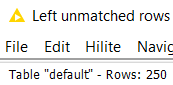
Ảnh có chứa bàn

Mô tả được tạo tự động

20)



21)

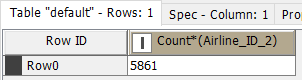


22)

Ảnh có chứa bàn

Mô tả được tạo tự động

23)



24)



=> Tuples deleted: 289 => 289/6150 = 4.699%

25)

In several previous task, we have the number of tuples removes:

Name: 2

Alias: 5

ITAT: 13

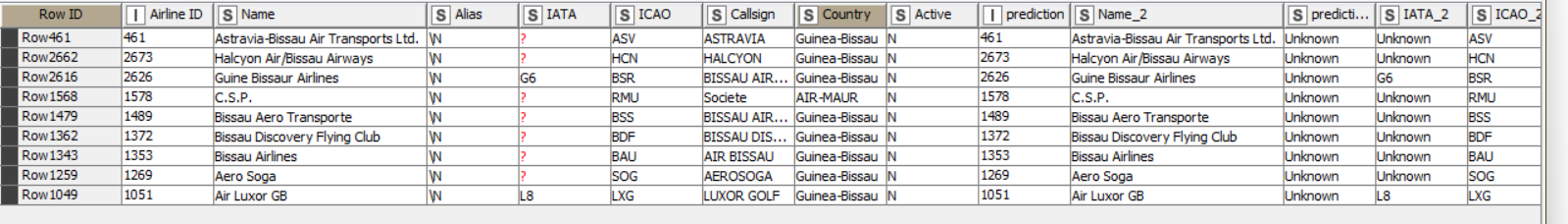
ICAO: 3

Callsign: 250

Country: 9

=> Callsign column has the most tuple being deleted

26)



2.

