**Assignment 1: Data Cleaning and Observation**

COS10022 – Introduction to Data Science

Semester 1, 2022

Names: Ngo Cong Thanh

Student ID: 103433609

Date: 06/02/ 2022

1. Observed Information:

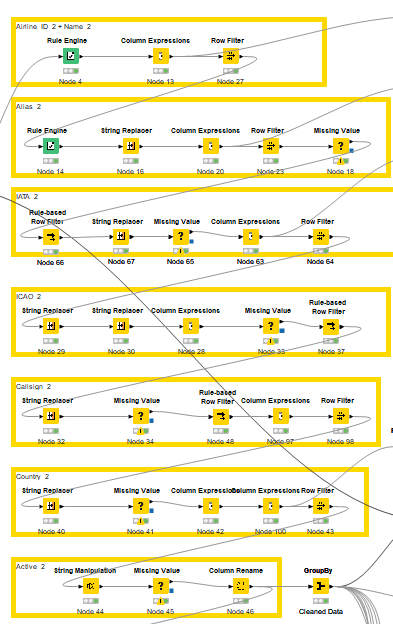
|  |  |  |  |
| --- | --- | --- | --- |
| **Question** | **Answer** | **Question** | **Answer** |
|  | 6150 |  | 5861 |
|  | 5778 |  | 164 |
|  | 1078 |  | 5566 |
|  | 5023 |  | 229 |
|  | 2 |  | 1 |
|  | 5698 |  | 4388 |
|  | 266 |  | 800 |
|  | 15 |  | 0 |
|  | 2 |  | 5 |
|  | 13 |  | 4 |
|  | 250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

1. [Put your answers here.]

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

3.1 Cleaning data process:

For this assignment, our team have read and consider about the solution of this data set. At first, we tried to take the whole data set and clean but it never came out as the best model to clean. So, we decided to tracking each column to meet its’ criteria.



Our idea of this model is very clear. We sort all invalid data (non-English letter or numerical, more than 2, more than 3 letters), and change them to a single character (n). After that, delete all that characters.

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

* 2.