**TEAM MEMBERS DETAILS**

GROUP NAME: PACHACUTEAM

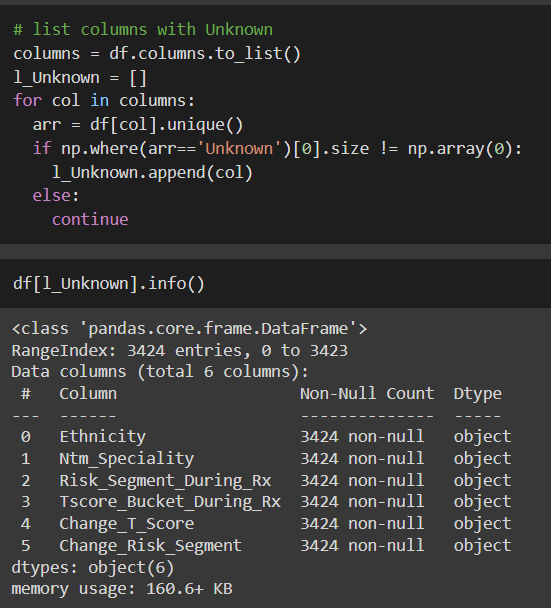
SPECIALIZATION: Data Science

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Email | Country | College/Company |
| Andersson Romero | [andromdez@gmail.com](mailto:andromdez@gmail.com) | Perú | Freelancer |

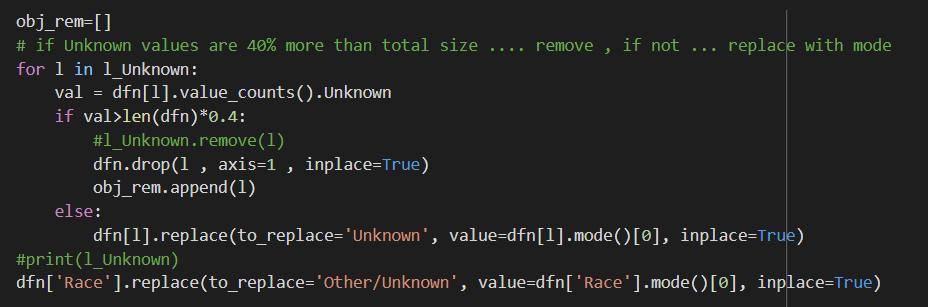
**Data Clean and Transform**

In this section we focus on cleaning outliers, replacing unknown data, and transforming categorical data.

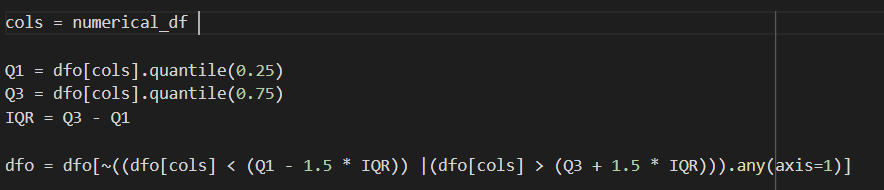
We start by looking to complete the unknown data, we first locate the features that contain unknown data.

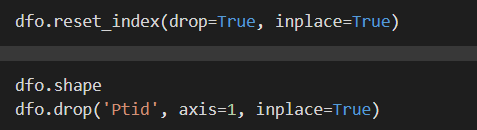


Once located, if these data exceed 40% of the total, the column is deleted. Otherwise, we replace with the mode (since they are categorical data).

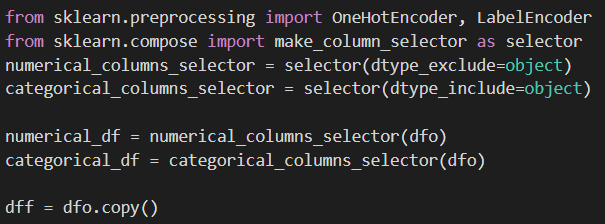


For the second part, we use quantiles to remove outliers. Additionally, remove them from the Ptid column.

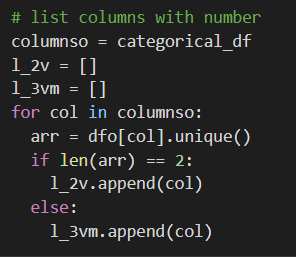




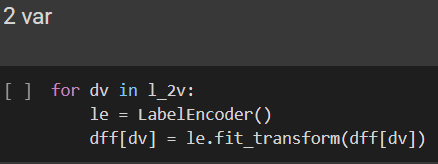
On the other hand, we convert the categorical variables to integers.

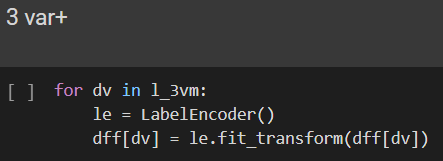


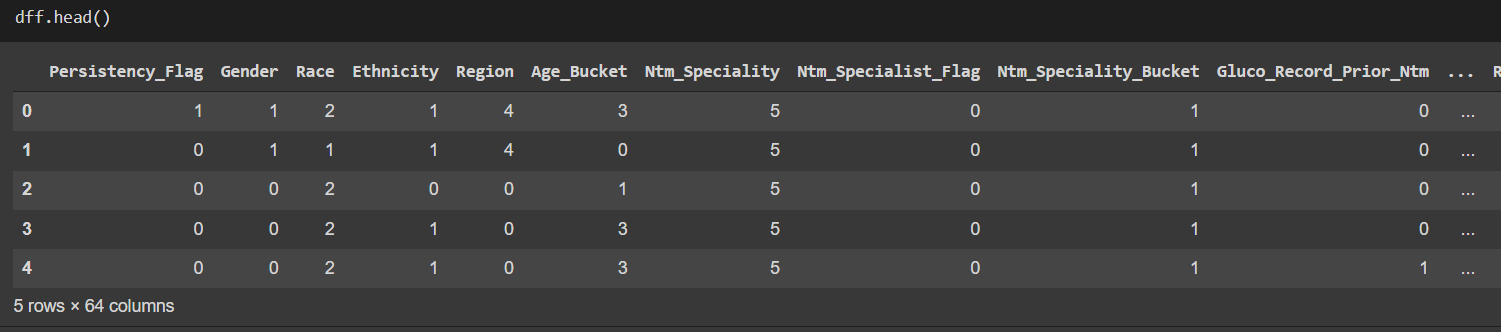
We first store the features with two categories and three or more categories.



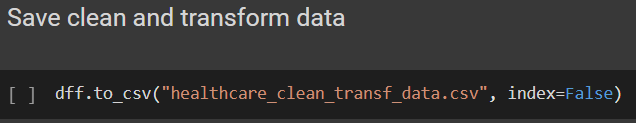
We convert, with LabelEncoder()







We save the clean and transformed dataset.



**DATA INTAKE REPORT**

Name: Healthcare project

Report date: 08/04/2022

Internship Batch: [LISUM07](https://canvas.instructure.com/courses/3110044)

Version: 1.0

Data intake by: PACHACUTEAM

Data intake reviewer:<intern who reviewed the report>

Data storage location: GitHub

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 3424 |
| **Total number of files** | 1 |
| **Total number of features** | 69 |
| **Base format of the file** | .xlsx |
| **Size of the data** | 899 KB |

**GITHUB LINK:** https://github.com/And2300/Healthcare\_PersistencyDrug