

DEC 2020

CLOUD COUNSELAGE

DATA CLEANING ALGORITHM

Version: 1

Created By: Sona Joseph

Created On: 28th Dec 2020

Approved By: Harish Sapte

Approved On:

.

**INDEX**

[1.Introduction 2](#_Toc59538792)

[2.Overview 3](#_Toc59538793)

[3. Methodology 4](#_Toc59538794)

[4.Conclusion 5](#_Toc59538795)

# 1.Introduction

Data Cleaning is the major task for everyone who is working on any kind of project related to data science or any Datasets.

Data cleaning consume the maximum time of data scientist. Due to which it may possible that he/she may lag with time for other necessary programming.

So, the POC is about the algorithm which may do major part of data cleaning in a minimal time. The algorithm will deal with the missing data. It will handle the duplicate and irrelevant data. Also this will perform converting datatypes.

So, data scientist will not have to waste their time in solving the common mistakes that each and every datasets may have.

# 2.Overview

[Write an overview to your Proof of Concept. Overview will contain the method you are trying to use to achieve a successful result]. (Arial 11)

For the implementation of data cleaning I have used pandas and numpy libraries. Following are the methods I have tried to implement.

* Handle missing data

In every datasets we may come across some entries to be missing. So, this may lead us to improper results.

Handling data can be done in two ways : 1) Droping the entire row.

2) Impute it with mean, median or mode

* Remove Duplicate Observations

Removing duplicate entries may result us more accurately.

For this I have used drop\_duplicates() method of pandas library

* Removal of unwanted observation

In a dataset it may contain many column which may or may not be useful for our project.

So, to work in an efficient way I have written a method to drop unwanted observation.

* Drop missing values

If any important data of a column is missing then most probably we select to drop that entire row.

For that we use drop[‘column\_name’] method.

* Imputing missing values

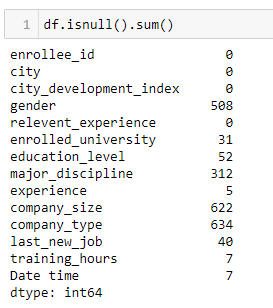
There are many ways to impute the missing values.

I have used mean(), median(), mode() method to impute the missing values in **int64** type of column.

# 3. Methodology

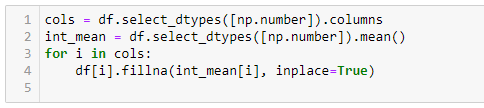
[Document the entire methodology in a step-wise manner. Add screenshots for every step if possible. Highlight the Screenshots for the main points]. (Arial 11)

For testing of algorithm I have taken a dataset. Below given is the total number of null values. Firstly we have to handle this missing data.

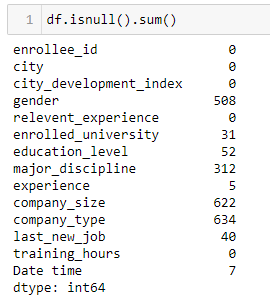


1. Handle missing data

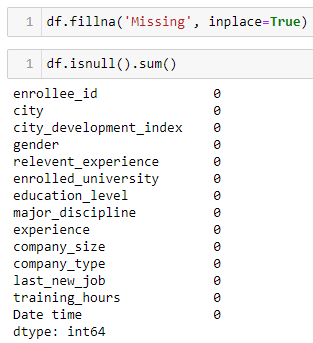
* Imputing mean to int type of data column



Output:

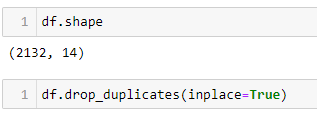


* For other datatypes imputing ‘missing’ in place of Nan.

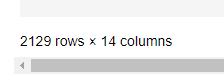


1. Dropping duplicates

* Dropping the duplicate entries

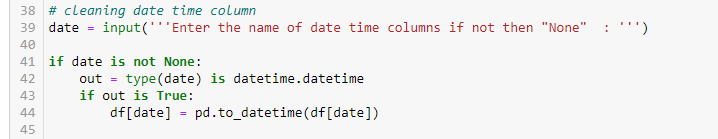


Output:



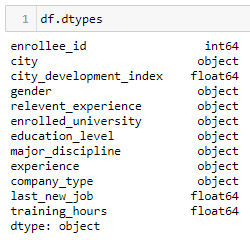
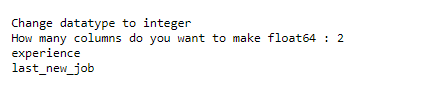
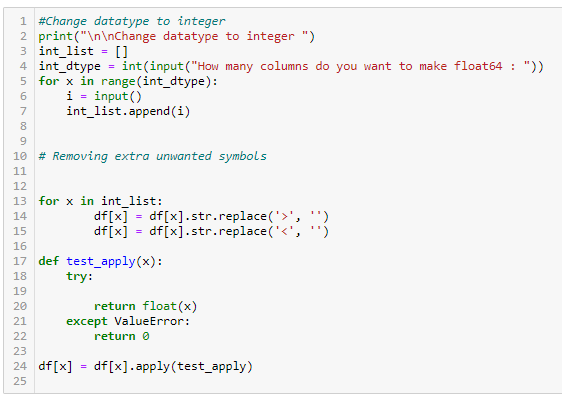
1. Handling date time columns.

* Cleaning date time

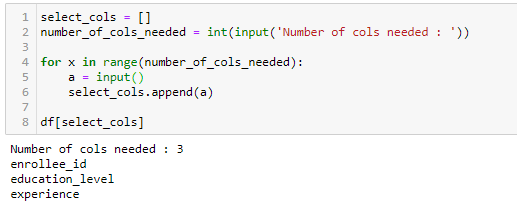


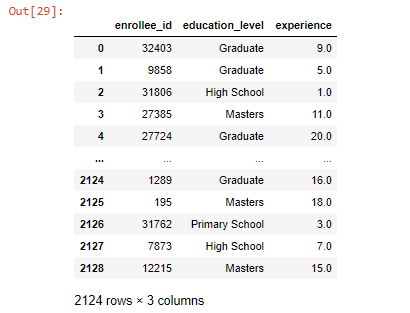
1. Changing datatype

* Converting column to float64

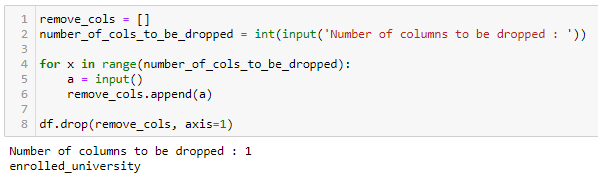


1. Selecting particular column

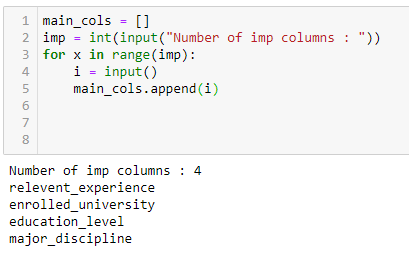




1. Dropping columns



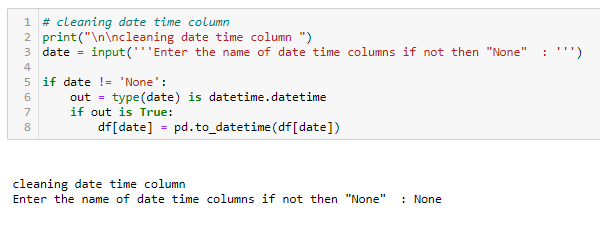
* Dropping Nan column of priority columns only.



Output:



1. Handling date time column error



# 4.Conclusion

[Draw the conclusions to your Proof of Concept and mention the reason for its success or failure]. (Arial 11)

POC : “We want to save the time of our employees who are working on solving Data Science problems. The idea here is to reduce the time required to perform some standardized steps to clean the data. To save the time of employees the data cleaning part will be done by an algorithm.”

In POC I have considered..

* Handle missing data
* Remove Duplicate Observations
* Remove Unwanted Observations
* Drop missing value data
* Impute missing data
* Change datatypes
* Cleaning date time column

This algorithm will priorly perform all the important steps to be performed for data cleaning. Like the methods mentioned above.

Main concept here is to save the time of data scientist. So, it is successful. But if the data scientist wants to do deep cleaning then he should do some more efforts that’s the feedback of my algorithm.