DATA WRANGLING

Dataset location : [incident\_managment\_process\_enriched\_event\_log](https://archive.ics.uci.edu/ml/datasets/Incident+management+process+enriched+event+log)

Jupyter Notebook : [github\_link](https://github.com/RahulKG7/Springboard_for_rahul/blob/master/Capstone_projet1/Corrected_project.ipynb)

Above are all the relevant files used in this project. Incident management process enriched event log data set is raw information in csv format that has been taken from UCI repository as mentioned above.

This dataset as mentioned in UCI website has missing values and it needs to be corrected before performing any other operations. Below are the finds from initial findings from the raw dataset.

* Data set consists of 141712 rows and 36 columns.
* There are multiple columns with missing values marked in the dataset by ‘?’.
* Also found 5 different columns with inaccurate data types.
* As a subject matter column “incident\_managment” one extra categorical value that was out of place, namely ‘-100’.

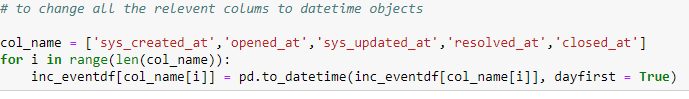
First course of action was to read the data.



Next step is to replace ‘?’ which symbolizes the missing values throughout the dataset with NaN for better maneuverability of data further on .



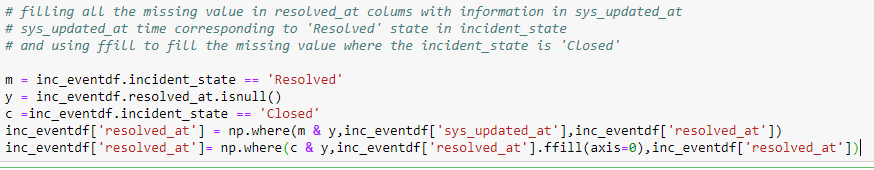
Moving forward to column data types we found earlier that there are 5 different columns was wrong data types. We solve the issue as shown below.



Next issue in our list was to correct information in the dataset. Such as the category “-100” in the “incident\_state” column. Which was replaced based on subject matter knowledge.



Further action is required to address missing values in the dependent variable column “resolved\_at” .



After performing all the above actions there are still some values that have missing entry, however they make up only 14 unique incidents out of 24918 hence are negligible and can be dropped.

